Differential abundance analysis for microbial marker-gene surveys

Joseph N Paulson[1,2], O Colin Stine[3], Héctor Corrada Bravo[1,2,4] & Mihai Pop[1,2,4]

We introduce a methodology to assess differential abundance in sparse high-throughput microbial marker-gene survey data. Our approach, implemented in the metagenomeSeq Bioconductor package, relies on a novel normalization technique and a statistical model that accounts for undersampling—a common feature of large-scale marker-gene studies. Using simulated data and several published microbiota data sets, we show that metagenomeSeq outperforms the tools currently used in this field.

Marker-gene surveys have recently been applied to clinical settings in order to understand the structure and function of healthy microbial communities and the association of microbiota with diseases such as Crohn's disease[1], bacterial vaginosis[2], diabetes[3], eczema[4], obesity[5] and periodontal disease[6]. Potentially pathogenic or probiotic bacteria can be identified by detecting significant differences in their distribution across healthy and disease populations, thereby making the analysis of differential abundance critical. Similar issues are encountered in the attempt to correlate microbiome composition with environmental factors. Although methods for comparing whole communities are commonly used in this context[7,8], there is a need for tools that discern taxon-specific associations in marker-gene surveys. We present a method that provides this level of resolution while removing biases that exist in current approaches.

The 16S ribosomal RNA gene is a commonly used marker for profiling diversity in microbial samples. Hypervariable regions within the gene are amplified and sequenced, and sequence reads are clustered into operational taxonomic units (OTUs)[9]. Representative sequences from each cluster are then classified taxonomically by alignment against a database of previously characterized 16S ribosomal DNA (rDNA) reference sequences[10].

Although data preprocessing and differential abundance analysis have been extensively studied in microarray and high-throughput sequencing (serial analysis of gene expression, or SAGE, and RNA-seq) experiments that measure gene expression, the specific characteristics of marker-gene data have required the development of specialized analytical tools[11–13]. The principal difference is that, unlike RNA-seq studies, most OTUs in marker-gene studies are rare (that is, absent from a large number of samples). This sparsity is due to both biological and technical phenomena: some organisms are found in only a small percentage of samples, whereas others are simply not detected owing to insufficient sequencing depth. These phenomena can lead to strong biases when data sets are scaled for comparison and when sequence read counts are tested for significant differences.

We present two complementary methods to analyze large-scale marker-gene microbial survey data, implemented in the publicly available metagenomeSeq Bioconductor package (http://cbcb.umd.edu/software/metagenomeSeq/). A normalization method avoids biases due to uneven sequencing depth, and a distribution mixture model removes testing biases caused by undersampling.

The most common normalization technique, total-sum scaling (TSS), divides feature read counts (the number of reads from a particular sample that cluster within the same OTU) by the total number of reads in each sample, i.e., it converts feature counts to appropriately scaled ratios. TSS has been shown to bias differential abundance estimates in RNA-seq data[14,15] because a few measurements (for example, taxa or genes) are sampled preferentially as sequencing yield increases, and these measurements therefore have an undue influence on normalized counts. An alternative is to scale across only the segment of the count distribution that is relatively invariant across samples. A recent proposal for normalization of RNA-seq data is to scale counts by the 75th percentile of each sample's nonzero count distribution[14]. The percentile that captures the relatively invariant count distribution varies across 16S rDNA data sets (**Supplementary Fig. 1**). Our cumulative-sum scaling (CSS) method is an adaptive extension of this approach that is better suited for marker-gene survey data. With CSS, raw counts are divided by the cumulative sum of counts up to a percentile determined using a data-driven approach.

We applied CSS normalization to data from a longitudinal study tracking the gut microbial community of twelve gnotobiotic mice[16]. To assess the effect of normalization on distinguishing samples by phenotypic similarity, we performed a multidimensional scaling analysis of data normalized using CSS, DESeq[17] size factors, trimmed mean of *M*-values[18] and total-sum normalization (**Fig. 1a–d**). CSS normalization was the best of these methods for separating samples according to diet while controlling within-group variance. We quantified this observation using linear

[1]Graduate Program in Applied Mathematics & Statistics, and Scientific Computation, University of Maryland, College Park, Maryland, USA. [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [3]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, Maryland, USA. [4]Department of Computer Science, University of Maryland, College Park, Maryland, USA. Correspondence should be addressed to M.P. (mpop@umiacs.umd.edu) or H.C.B. (hcorrada@umiacs.umd.edu).

RECEIVED 4 MARCH; ACCEPTED 16 AUGUST; PUBLISHED ONLINE 29 SEPTEMBER 2013; DOI:10.1038/NMETH.2658

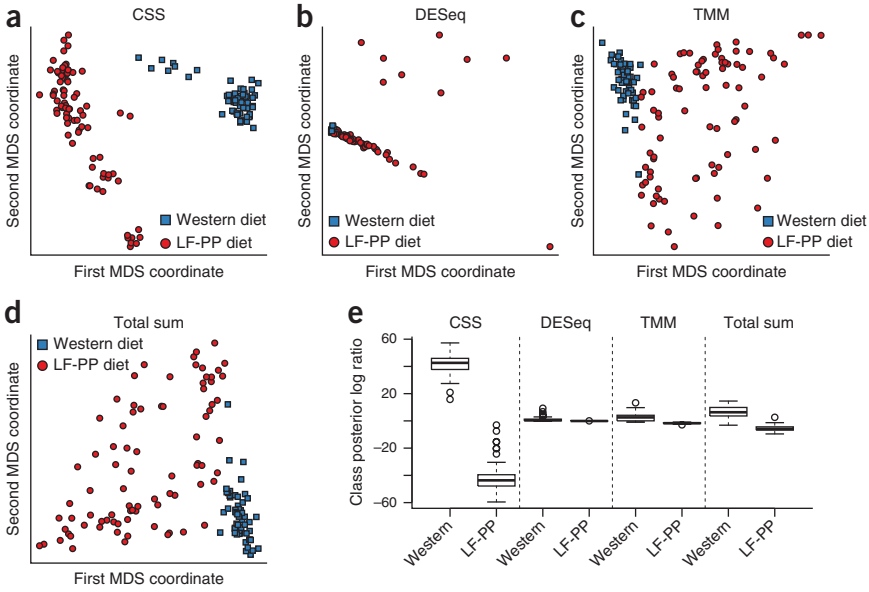

Figure 1 | Clustering analysis is improved substantially by CSS normalization. (**a**–**d**) We plot the first two principal coordinates in a multidimensional scaling (MDS) analysis of mouse stool data normalized by CSS (**a**), DESeq size factors (**b**), trimmed mean of *M*-values (TMM) (**c**) and total-sum scaling (**d**). Colors indicate clinical phenotype (diet). LF-PP, low-fat, plant polysaccharide–rich diet. CSS normalization of data successfully separates samples by diet while controlling within-group variability. (**e**) Class posterior probability log-ratio for Western diet obtained from linear discriminant analysis. Each box corresponds to the distribution of leave-one-out posterior probability of assignment to the "Western" cluster across normalization methods (whiskers indicate 1.5× interquartile range). Samples were best distinguished by phenotypic similarity using CSS normalization.

discriminant analysis (Online Methods), confirming that CSS normalization best distinguished samples by phenotypic similarity (**Fig. 1e**). Comparison to other frequently used normalization methods gave similar results (**Supplementary Fig. 2**).

Our second contribution is a zero-inflated Gaussian (ZIG) distribution mixture model that accounts for biases in differential abundance testing resulting from undersampling of the microbial community. We found a strong correlation between the number of OTUs detected in a sample and the corresponding sequencing depth in high-throughput 16S rDNA studies (R^2 = 0.92–0.97; **Supplementary Fig. 3**), which is consistent with previous reports[19–21]. This suggests that measurements of differential abundance suffer from biases because zero counts in samples with low coverage are misinterpreted as absent taxonomic features when they are in fact the result of undersampling. The degree of sparsity observed in marker-gene experiments (1–3% genes detected) is much higher than that usually seen in other abundance assays such as transcriptome profiling from single genomes (15–85% genes detected; **Supplementary Fig. 4**).

To explicitly account for undersampling, we use a mixture model that implements a ZIG distribution of mean group abundance for each taxonomic feature (**Supplementary Fig. 5**). The effect of this model is exemplified on one OTU from the Human Microbiome Project[22] (**Supplementary Fig. 6**). Using posterior probability estimates that account for community undersampling as weights to estimate count distribution parameters reduced the estimated fold change between the two groups under study. Furthermore, counts after accounting for undersampling were better fit by a log-normal distribution (Shapiro-Wilks test, P = 0.78) than were normalized counts (Shapiro-Wilks test, P = 0.08).

We evaluated metagenomeSeq using simulated data and compared the results to those of the metagenomic tools Metastats[12], Xipe[11] and a Kruskal-Wallis test as used in LEfSe (linear discriminant analysis effect size)[13] as well as representative methods for RNA-seq analysis. metagenomeSeq and, to a lesser degree, the Kruskal-Wallis test consistently produced high area under the curve (AUC) scores across most simulation settings (**Fig. 2**). However, metagenomeSeq obtained the highest AUC of these methods (including LEfSe's Kruskal-Wallis test) in data sets with high sparsity, which actual metagenomic data sets tend to have (>85% sparsity; **Fig. 2a**). Metastats, edgeR[18] and DESeq[17] had similar performance characteristics to each other, with smaller AUC scores. Xipe performed poorly across most simulation settings, as expected, because this method does not account for population variability.

Our ZIG model uses linear modeling following standard conventions in methods for testing differential abundance in gene expression[23] that control for confounding factors. In contrast, LEfSe uses an *ad hoc* heuristic approach to account for subpopulations in large marker studies that is overly conservative and prone to low sensitivity. We observed by simulation (**Supplementary Fig. 7** and **Supplementary Table 1**) that metagenomeSeq was more sensitive than LEfSe (0.95 vs. 0.01, respectively) and retained high specificity (0.96 vs. 1) when confounding subpopulations were included among tested groups.

We also compared these methods using oral microbiota data from the Human Microbiome Project[22] (**Supplementary Fig. 8**) to identify OTUs that are differentially abundant between tongue and subgingival plaque samples. Metastats and edgeR identified the largest number of significant OTUs (533 and 524, respectively), whereas metagenomeSeq (360) and, especially, DESeq (20) and LEfSe (8) identified many fewer significant OTUs (**Supplementary Table 2**). Organisms found enriched in subgingival plaque by metagenomeSeq but missed by DESeq or LEfSe are fairly abundant, well-known members of the periodontal microbiome and include sulfate-reducing bacteria, which have been proposed as potential pathogenicity factors in periodontal disease[24]. In general, the poor performance of Metastats and LEfSe was due to their lack of robust modeling of confounding factors; for DESeq and edgeR, the assumptions upon which these models are based were not met by these data. We provide a detailed comparison of these results in the **Supplementary Note** and **Supplementary Figures 9–11**.

We further compared our results at the species level with those obtained by Segata *et al.*[25], who applied LEfSe to the same oral data set. While we confirmed all species detected as differentially abundant by LEfSe, we also identified with metagenomeSeq three

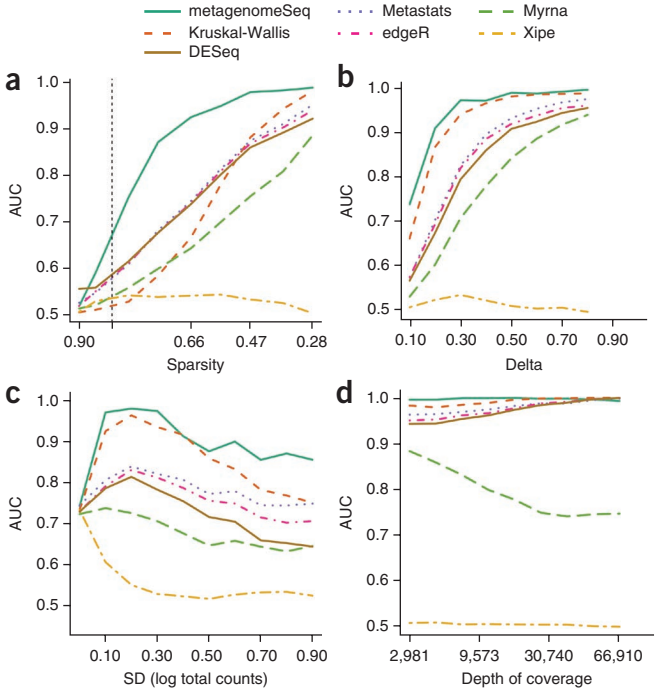

Figure 2 | Simulation results indicated that metagenomeSeq has greater sensitivity and specificity in a variety of settings. We used the area under the receiver operating characteristic curve (AUC) to compare Metastats[12], Xipe[11], the Kruskal-Wallis test as used in LEfSe[13], a nonzero inflated log-normal model (Myrna)[28], edgeR[18] and DESeq[17]. (**a**) AUC as data set sparsity decreases. metagenomeSeq achieved larger AUC values than any other method in data sets with high sparsity (vertical dashed line represents the least sparse metagenomic data set). (**b**) AUC as the effect size between two conditions (Delta) increases. Both metagenomeSeq and LEfSe were better at detecting features with small effect size. (**c**) AUC as the variability in depth of sequencing increases. metagenomeSeq and the Kruskal-Wallis test were robust with respect to high variability in sequencing depth. SD, sampling depth s.d. (**d**) AUC as average sequencing depth increases. All models (except the nonzero inflated log-normal model and Xipe) performed similarly well at sufficient depth of coverage.

additional species missed by their analysis. Specifically, we found *Atopobium parvulum*, *Lautropia* sp. and *Desulfotomaculum* sp. to be enriched in subgingival plaque (**Supplementary Fig. 12**). All of these were fairly abundant in the samples (at least 4% of the population) and represent previously characterized members of the normal subgingival microbiota[24,26,27].

In summary, our methods yield a more precise biological interpretation of the data. In mouse stool data, CSS normalization helps distinguish clinical phenotypes that are confounded by commonly used normalization methods, whereas in the oral microbiome, the combined differential abundance modeling approach identifies associations that were missed by commonly used tools. Although undersampling is ubiquitous in marker-gene survey data, to our knowledge, the approach presented here is the first to correct for this phenomenon. Though our focus is on data generated in microbial community surveys, sparsity may also be an issue in some RNA-seq experiments, and thus our methods may have broader applicability (**Fig. 2a**).

This work addresses some of the main challenges to robust analysis of marker-gene surveys in clinical and epidemiological settings: (i) variable depth of coverage across samples and the resulting rarefaction effect and (ii) confounding due to technical and population characteristics. We expect that metagenomeSeq will help practitioners achieve the full promise of marker-gene surveys in clinical research.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

J.N.P. was supported by a US National Science Foundation Graduate Research Fellowship (award DGE0750616). J.N.P., O.C.S. and M.P. were supported in part by the Bill and Melinda Gates Foundation (award 42917 to O.C.S.). H.C.B. was supported in part by the US National Institutes of Health grant 5R01HG005220. We would like to thank B. Lindsay and L. Magder for discussion of the methods and C.M. Hill for help with clustering of OTUs.

AUTHOR CONTRIBUTIONS

J.N.P. and H.C.B. developed the algorithms and wrote the software. J.N.P. collected results. O.C.S. and M.P. contributed to discussions of the methods. J.N.P., H.C.B. and M.P. analyzed results. J.N.P., H.C.B. and M.P. wrote the manuscript. All authors read and approved the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Morgan, X.C. *et al. Genome Biol.* **13**, R79 (2012).
2. Ravel, J. *et al. Proc. Natl. Acad. Sci. USA* **108** (suppl. 1), 4680–4687 (2011).
3. Qin, J. *et al. Nature* **490**, 55–60 (2012).
4. Harris, J.K. & Wagner, B.D. *J. Allergy Clin. Immunol.* **129**, 441–442 (2012).
5. Turnbaugh, P.J. *et al. Nature* **457**, 480–484 (2009).
6. Scher, J.U. *et al. Arthritis Rheum.* **64**, 3083–3094 (2012).
7. Caporaso, J.G. *et al. Nat. Methods* **7**, 335–336 (2010).
8. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
9. Ghodsi, M., Liu, B. & Pop, M. *BMC Bioinformatics* **12**, 271 (2011).
10. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).
11. Rodriguez-Brito, B., Rohwer, F. & Edwards, R.A. *BMC Bioinformatics* **7**, 162 (2006).
12. White, J.R., Nagarajan, N. & Pop, M. *PLoS Comput. Biol.* **5**, e1000352 (2009).
13. Segata, N. *et al. Genome Biol.* **12**, R60 (2011).
14. Bullard, J.H., Purdom, E., Hansen, K.D. & Dudoit, S. *BMC Bioinformatics* **11**, 94 (2010).
15. Dillies, M.A. *et al. Brief. Bioinform.* doi:10.1093/bib/bbs046 (17 September 2012).
16. Turnbaugh, P.J. *et al. Sci. Transl. Med.* **1**, 6ra14 (2009).
17. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
18. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. *Bioinformatics* **26**, 139–140 (2010).
19. White, J.R. *et al. BMC Bioinformatics* **11**, 152 (2010).
20. Friedman, J. & Alm, E.J. *PLoS Comput. Biol.* **8**, e1002687 (2012).
21. Faust, K. *et al. PLoS Comput. Biol.* **8**, e1002606 (2012).
22. The Human Microbiome Project Consortium. *Nature* **486**, 215–221 (2012).
23. Smyth, G.K. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S.) 397–420 (Springer, 2005).
24. Langendijk-Genevaux, P.S., Grimm, W.D. & van der Hoeven, J.S. *J. Clin. Periodontol.* **28**, 1151–1157 (2001).
25. Segata, N. *et al. Genome Biol.* **13**, R42 (2012).
26. Paster, B.J. *et al. J. Bacteriol.* **183**, 3770–3783 (2001).
27. Colombo, A.P. *et al. J. Periodontol.* **80**, 1421–1432 (2009).
28. Langmead, B., Hansen, K.D. & Leek, J.T. *Genome Biol.* **11**, R83 (2010).

ONLINE METHODS

Cumulative-sum scaling normalization. We assume raw data are given as count matrix $M(m, n)$, where m and n are the number of features and samples, respectively. The raw data in this matrix are represented by counts $c_{i,j}$, the number of times taxonomic feature i was observed in sample j. We denote the sum of counts for sample i as $s_j = \Sigma_i c_{ij}$. The usual normalization procedure for marker-gene survey data corresponds to producing normalized counts $\widetilde{c_{ij}} = c_{ij} / s_j$. We refer to this procedure as total-sum normalization.

We introduce a new normalization method, cumulative-sum scaling normalization (CSS), to remove biases in the count data. The biases come from features that are preferentially amplified in a sample-specific manner. We denote the lth quantile of sample j as q_j^l, that is, in sample j there are l taxonomic features with counts smaller than q_j^l. For $l = 0.95m$, q_j^l corresponds to the 95th percentile of the count distribution for sample j.

We also denote

$$s_j^l = \sum_{i|c_{ij} \le q_j^l} c_{ij}$$

as the sum of counts for sample j up to the lth quantile. Using this notation, the total sum $s_j = s_j^m$. Our normalization chooses a value $\hat{l} \le m$ to define a normalization scaling factor for each sample to produce normalized counts $\widetilde{c_{ij}} = (c_{ij} / s_j^{\hat{l}})(N)$, where N is an appropriately chosen normalization constant. We scale all samples using the same constant N so normalized counts have interpretable units. We recommend using the median scaling factor $s_j^{\hat{l}}$ across samples. Counts for samples with a scaling factor close to N can be interpreted as reference samples, and counts for other samples are interpreted relative to the reference. In our data sets the median $s_j^{\hat{l}}$ was close to 1,000, and we thus used this value in our analysis. Note that ratios are also used in this procedure, assuming there is a finite capacity to the size of microbial communities. This is the same assumption that underlies total-sum normalization. However, our method seeks to avoid placing undue influence on features that are preferentially sampled. The relative proportion of the features is unaffected by the normalization as $s_j = \Sigma_i c_{ij}$ and $\tilde{s}_j = \Sigma_i c_{ij} / s_j^{\hat{l}}$. This implies

$$p_i = \frac{c_{ij}}{s_j} = \frac{s_j^{\hat{l}} \times c_{ij}}{s_j^{\hat{l}} \times \sum_i c_{ij}} = \frac{\widetilde{c_{ij}}}{\tilde{s}_j} = \tilde{p}_i$$

The choice of the appropriate quantile given by $\hat{l}$ above is critical for ensuring that the normalization approach does not introduce normalization-related artifacts in the data. At a high level, the count distribution of samples should all be roughly equivalent and independent of each other up to this quantile under the assumption that, at this range, counts are derived from a common distribution. The specific value for the chosen quantile is project specific and likely depends on the complete experimental details (including all the sample preparation, sequencing, and subsequent bioinformatics analysis).

We use an adaptive, data-driven method to determine $\hat{l}$ based on the observation above. We find a value $\hat{l}$ where sample-specific count distributions deviate from an appropriately defined reference distribution. Specifically, denote $\bar{q}^l = \text{med}_j\{q_j^l\}$, the median lth quantile across samples, as the lth quantile of the reference distribution. Note that this is exactly the way a reference distribution is defined in the commonly used quantile normalization approach[15]. We denote as $d_l = \text{med}_j\,|q_j^l - \bar{q}^l|$ the median absolute deviation of sample-specific quantiles around the reference. Under the methods assumptions, this quantity d_l is stable for low quantiles and shows high instability in high quantiles. Our method defines $\hat{l}$ as the smallest value for which high instability is detected (**Supplementary Fig. 1**). We measure instability in this case by using relative first differences. Specifically, we set $\hat{l}$ to the smallest l that satisfies $d^{l+1} - d^l \ge 0.1d^l$. The value 0.1 is set arbitrarily and may be substituted by another value to determine high instability.

We found that CSS-normalized sample abundance measurements are well approximated by a log-normal distribution in studies with large numbers of samples; we therefore applied a logarithmic transform to the normalized count data. This transformation controls the variability of taxonomic feature measurements across samples (**Supplementary Fig. 13**).

Assessment of normalization methods. To assess the effect of normalization on distinguishing samples by phenotype, we performed a multidimensional scaling analysis of count data normalized using CSS, total-sum scaling, logged total-sum scaling, geometric mean, trimmed mean by M-values, quantile scaling, and quantile normalization.

We calculated the 1,000 taxonomic features with the largest variance after each normalization method and used those normalized feature counts in the MDS analysis. We also used linear discriminant analysis (LDA) to distinguish samples by diet. We calculated the log ratio of class posterior probabilities for each sample x using leave-one-out cross-validation

$$\log \frac{f_w(x)\pi_w}{f_l(x)(1-\pi_w)}$$

where π_w is the proportion of samples on the Western diet, and f_w and f_l are normal densities for each of the diets, with a common variance. Parameters in each leave-one-out fold are estimated from the remaining samples. The class posterior probability should be large and positive for "Western" samples and small and negative for samples in the other group. We measure the performance of each normalization method by the difference in the distribution of the class posterior probabilities (**Fig. 1e** and **Supplementary Fig. 2e**).

Zero-inflated Gaussian Model. Our zero-inflated Gaussian (ZIG) mixture model is motivated by the observed relationship between depth of coverage and the number of OTUs detected (**Supplementary Fig. 3**). The components of the mixture model correspond to normally distributed log abundances in each group of interest: for example, case or control (represented as the count distribution in **Supplementary Fig. 5**) and a spike-mass at 0 indicating absence of the feature owing to undersampling (represented as the detection distribution in **Supplementary Fig. 5**). Our model seeks to directly estimate the probability that an observed zero is generated from the detection distribution owing to undersampling or from the count distribution (absence of the taxonomic feature in the microbial community). We estimate the expected value of latent component indicators on the basis of sample sequencing depth of coverage using an

expectation-maximization algorithm. A detailed description of the model is available in the **Supplementary Note**.

Simulation study. We simulated OTU-level data sets with 1,000 features. A sample's total count was sampled from a log-normal distribution with $\mu = 7.5$ and s.d. = 0.3. These values represent similar total counts to those observed in data. The first 50 features were chosen to be 'significant'. In one of the populations, for the first 25 significant features, we changed the proportion of the total counts for those features by adding $10^{-2} \times \delta$ percentage of the particular sample's total counts. For the remaining 25 we subtracted $10^{-2} \times \delta$ percentage of the sample's total counts. We used a logistic regression model of the proportion of zeros as a function of depth of coverage in a standard marker-gene survey to build a plausible simulation model for sparsity. Given a sample's depth of coverage s_j, an expected proportion of zero features π_j is obtained from the logistic regression fit. For each feature we randomly drew from a Bernoulli trial with probability π_j to spuriously set the feature to zero. Finally, we assigned randomly to 5% of the data an additional 1.3% (a value obtained from a standard marker-gene survey) of the mean of the total counts to introduce extremely abundant features.

Subgroup simulation. We simulated data from two populations, each of which consisted of two subpopulations. This example represents a case-control study in which cases and controls were collected from differing sites. We simulated OTU-level data sets with 1,000 features. A sample's total count was sampled from a log-normal distribution with $\mu = 7.5$ and s.d. = 0.3. These values represent similar total counts to those observed in data. The first 50 features were chosen to be significant. In one of the populations, for the first 25 significant features, we changed the proportion of the total counts for those features by adding $10^{-2} \times \delta$ percentage of the particular sample's total counts. For the remaining 25 we subtracted $10^{-2} \times \delta$ percentage of the sample's total counts. The second subgroup had a relatively larger expression of the significant features. This represents potential greater feature enrichment in a site's subpopulation. The trend, though, across populations in either subgroup is to either increase or decrease in cases or controls. Finally, 5% of the data are randomly given an additional 1.3% (a value obtained from a standard marker-gene survey) of the mean of the total counts to introduce extremely abundant features.

Marker-gene survey data. *Humanized gnotobiotic mouse gut.* Twelve germ-free adult male C57BL/6J mice were fed a low-fat, plant polysaccharide–rich diet. Each mouse was gavaged with healthy adult human fecal material. Following the fecal transplant, mice remained on the low-fat, plant polysaccharide–rich diet for 4 weeks, following which a subset of six mice were switched to a high-fat and high-sugar diet for 8 weeks. Fecal samples for each mouse went through PCR amplification of the bacterial 16S rRNA gene V2 region weekly. Details of experimental protocols and further details of the data can be found in Turnbaugh *et al.*[16] OTUs were classified by RDP[10] and annotated (minimum confidence level of 0.8). Sequences can be found at http://gordonlab.wustl.edu/TurnbaughSE_10_09/STM_2009.html.

Subgingival plaque and tongue dorsum. Subgingival plaque and tongue dorsum samples were a part of the Human Microbiome Project (HMP)[22] data set used in this analysis. The samples were part of a larger study aimed at cataloging the healthy human microbiome. Reads were deposited into the Data Analysis and Coordination Center (DACC). In particular, reads and metadata were downloaded from http://www.hmpdacc.org/HMR16S/. Further information on the data collection protocol and samples is available at http://www.hmpdacc.org/ and in the HMP paper[22]. Only patients from their earliest visit were considered, as were only samples that were properly annotated. Following OTU propagation (described below), singletons (up to five positive samples) were trimmed. To consider solely differential abundance estimates, we report on OTUs present in at least approximately 2% of the population. For each differential abundance method compared, differentially abundant OTUs were determined at FDR < 0.05, where the OTU is at least twice as abundant in one group compared to in the other (absolute estimated fold change >1). We used LEfSe's default detection method (as no fold change estimate is provided).

Human Microbiome Project data. Data used in **Supplementary Figure 3** were part of the HMP[22] data set used in this analysis. The samples were a catalog of the healthy human microbiome. Reads were organized into OTUs by QIIME[8] and deposited in the DACC. In particular, OTUs and metadata were downloaded from http://www.hmpdacc.org/HMQCP/. Further information on data collection protocol and samples is available at http://www.hmpdacc.org/ and in the HMP paper[22].

Lung microbiome. The lung microbiome consisted of respiratory flora sampled from six healthy individuals: three healthy nonsmokers and three healthy smokers. The upper lung tracts were sampled by oral wash and oro-nasopharyngeal swabs. Up to a patient's glottis, samples were taken using two bronchoscopes, a serial bronchoalveolar lavage, and lower-airway protected brushes. More detailed information about the lung microbiome samples, collection, and protocols is available in Charlson *et al.*[29] Reads and barcodes were provided by F. Bushman. Following OTU propagation (described below), OTUs were trimmed if they were not present in approximately 8% of the population.

Analysis pipeline. *OTU identification and annotation.* 454 SFF files and barcode dictionaries were downloaded and run through the same pipeline. Conservative OTUs were constructed by pooling together the sequences from all samples and were then clustered using DNAClust[9] with default parameters (99% identity clusters) to ensure that the definition of an OTU was consistent across all samples. To obtain taxonomic identification, a representative sequence from each OTU was aligned to data in the Ribosomal Database Project (RDP, http://rdp.cme.msu.edu/, release 10.4) using Blastn with long word length (-W 100) in order to detect nearly identical sequences only. Sequences without a nearly identical match to RDP were marked as having "no match" and assigned an OTU identifier. The resulting data were organized into a collection of tables at many different taxonomic levels containing each taxonomic group as a row and each sample as a column. These tables formed the substrate for the statistical analyses described. This process was performed for the HMP and the human lung microbiome data sets. After removing OTUs present in fewer than five samples, the HMP data set consisted of 23,685 OTUs, whereas the human lung microbiome consisted of 2,365 OTUs. We explored the effect of ambiguously

doi:10.1038/nmeth.2658

assigned reads (sequences that have good matches to two or more OTUs) by running DNAClust in "non-overlapping" mode: a mode that ensures high separation between clusters and eliminates ambiguous reads. We also ran the HMP data set using this option and confirmed all results shown in the paper (**Supplementary Figs. 14** and **15**). We provide further discussion of the ambiguity of mapping reads to OTUs in the **Supplementary Note**.

RNA-seq data. RNA sequencing counts were downloaded from ReCount[30], http://bowtie-bio.sourceforge.net/recount/. Only data sets with at least 15 samples were considered.

Software. The following software versions were used for analysis: DESeq v.1.8.3 (ref. 17), edgeR v.2.6.12 (ref. 18), and limma v.3.12.3 (ref. 23) were used in the comparisons. Personalized R scripts were written for the other methods, and all analyses were performed on R v.2.15.1 on a Red Hat Enterprise Linux Server release 5.9 (Tikanga) 64-bit platform.

29. Charlson, E.S. *et al. Am. J. Respir. Crit. Care Med.* **184**, 957–963 (2011).
30. Frazee, A.C., Langmead, B. & Leek, J.T. *BMC Bioinformatics* **12**, 449 (2011).