# Supporting Information

## Jeraldo et al. 10.1073/pnas.1206721109
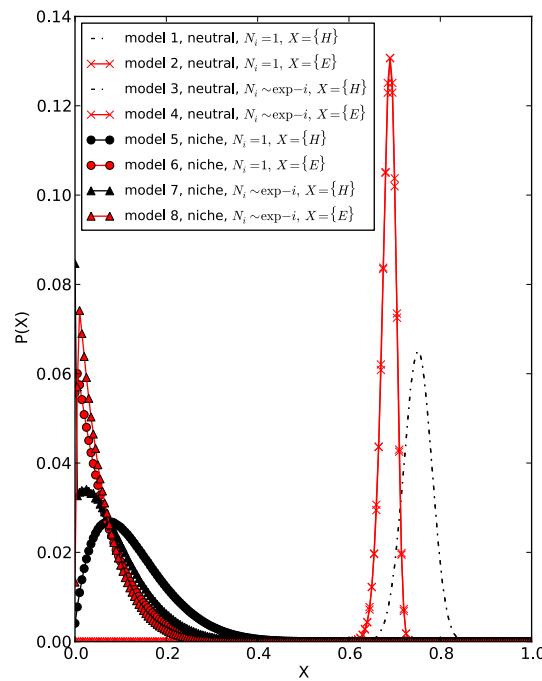


**Fig. S1.** Explicit numerical calculations of our metric on eight model systems. In these systems, we study the difference between the effects of the metric on neutral (models 1–4) and niche model systems (models 5–8). We also study the effect of choosing the closest distance (even-numbered models) compared with considering all distances (odd-numbered models). Finally, we consider the weighted models (models 3–4 and 7–8) vs. the unweighted ones (models 1–2 and 5–6).
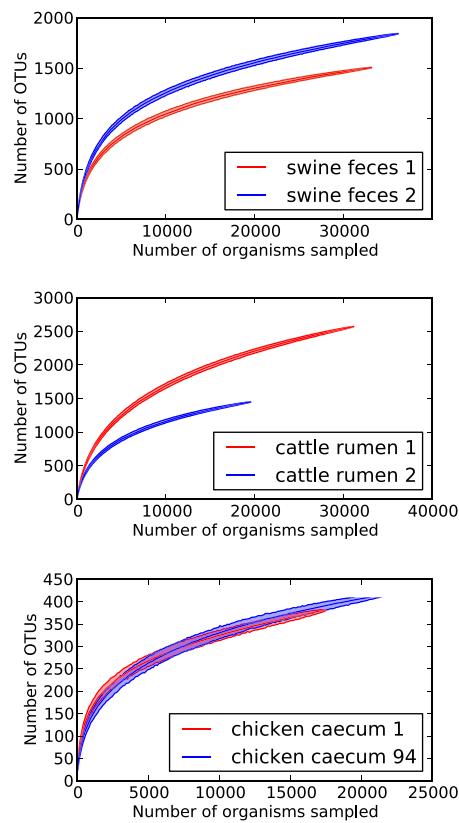
**Fig. S2.** Rarefaction curves for the six vertebrate GI microbiomes. The solid line represents the median number of OTUs (100 resamplings), whereas the shaded area represents the 95% confidence interval.
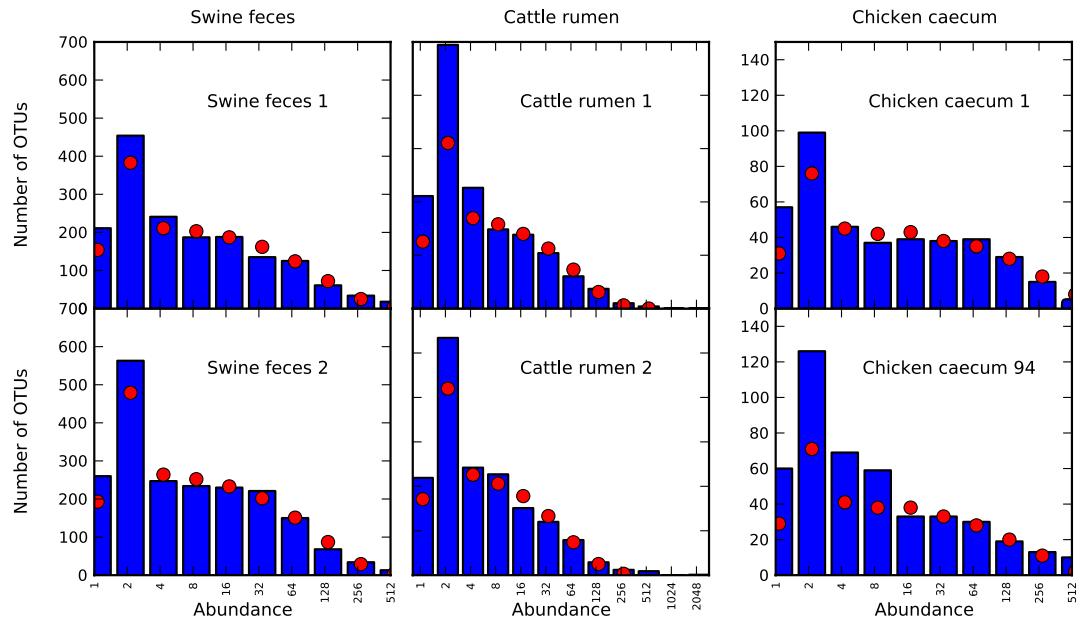


**Fig. S3.** Preston plot for swine feces, cattle rumen, and chicken ceca samples. In a Preston plot, the height of the bar indicates the number of species observed with abundance 1, 1–2, 2–4, and 4–8, for example. Note that in all six datasets, most OTUs are singletons. In this plot, one to two bars are highest because of an artifact. Traditionally, in a Preston plot, the OTUs with borderline abundances split evenly between two neighboring bins.
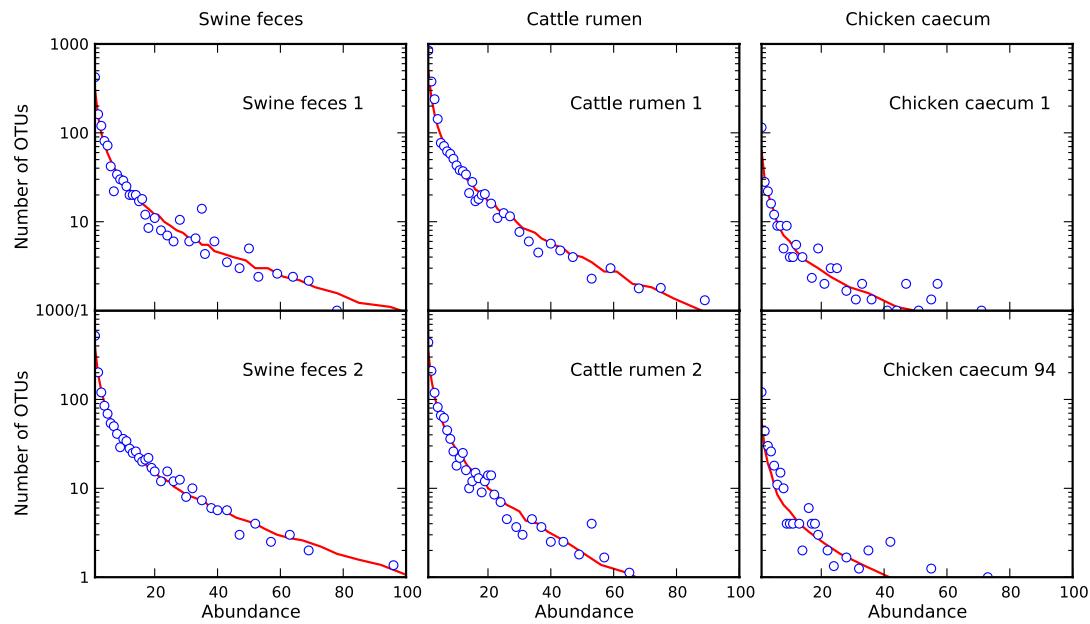
**Fig. S4.** Species abundance distribution for swine feces, cattle rumen, and chicken ceca. The species abundance distribution indicates the number of OTUs collected for each abundance.
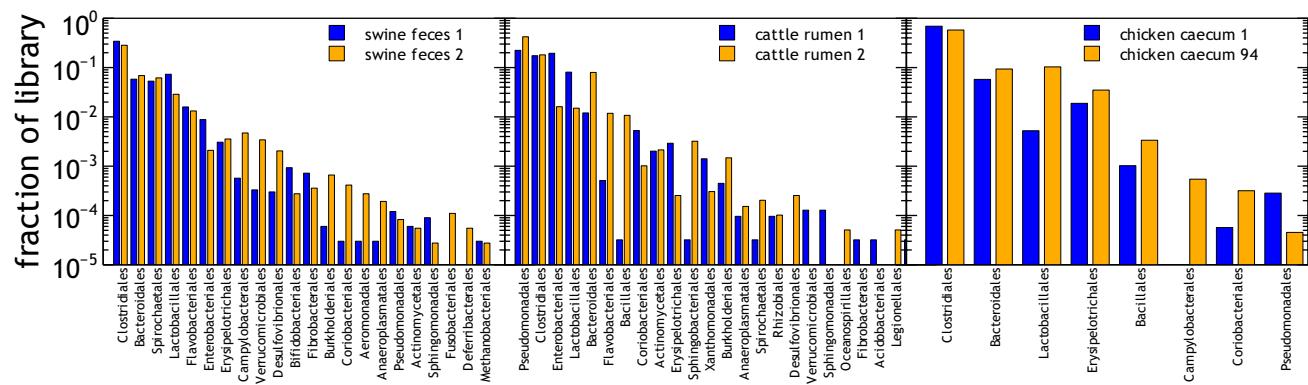


**Fig. S5.** Taxa comparisons. Taxonomic assignments at order level for all libraries, at 80% confidence threshold, sorted by combined abundance. Although there appear to be no differences in the form of the rank-abundance curves, we see differences in the taxonomic distributions here as a result of changes in diet or challenges to the microbial ecosystem.
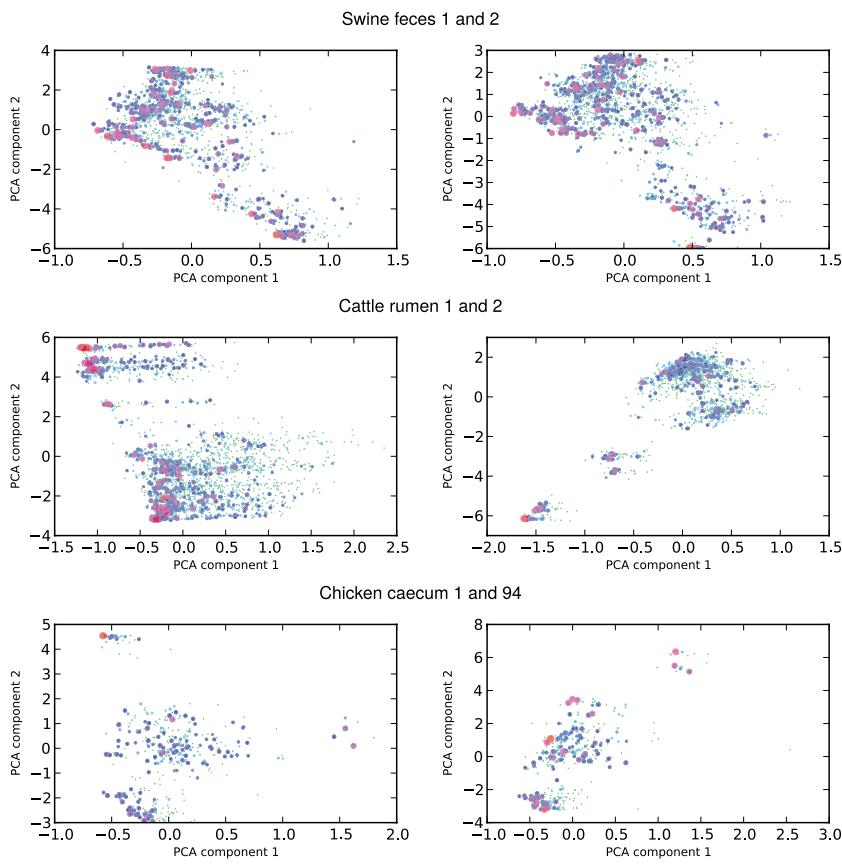
**Fig. S6.** Weighted PCA ordination applied to the six experimental datasets. Details on how weighted PCA was performed are provided in the main text. Each circle in this figure represents an OTU, and its size and color indicate the logarithm of OTU abundance.
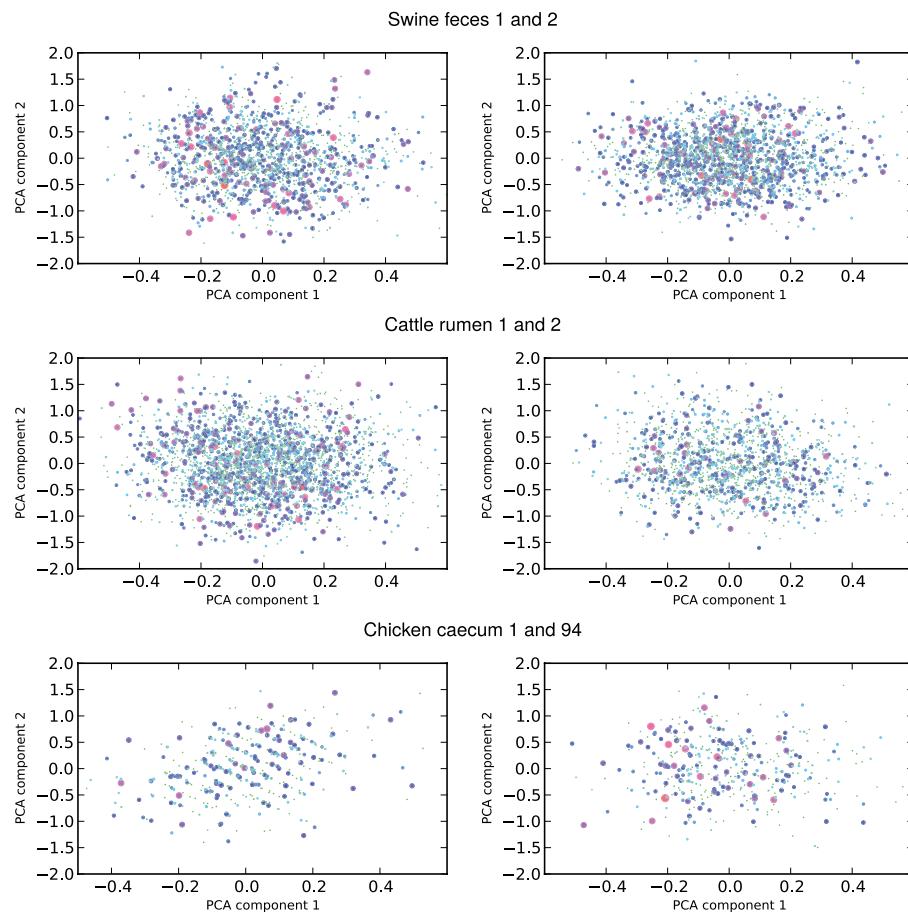
**Fig. S7.** Weighted PCA ordination applied to the randomized datasets (compare with Fig. S5). Details on how the randomized datasets were generated and how weighted PCA was performed are provided in the main text. Each circle in this figure represents an OTU, and its size and color indicate the logarithm of OTU abundance.
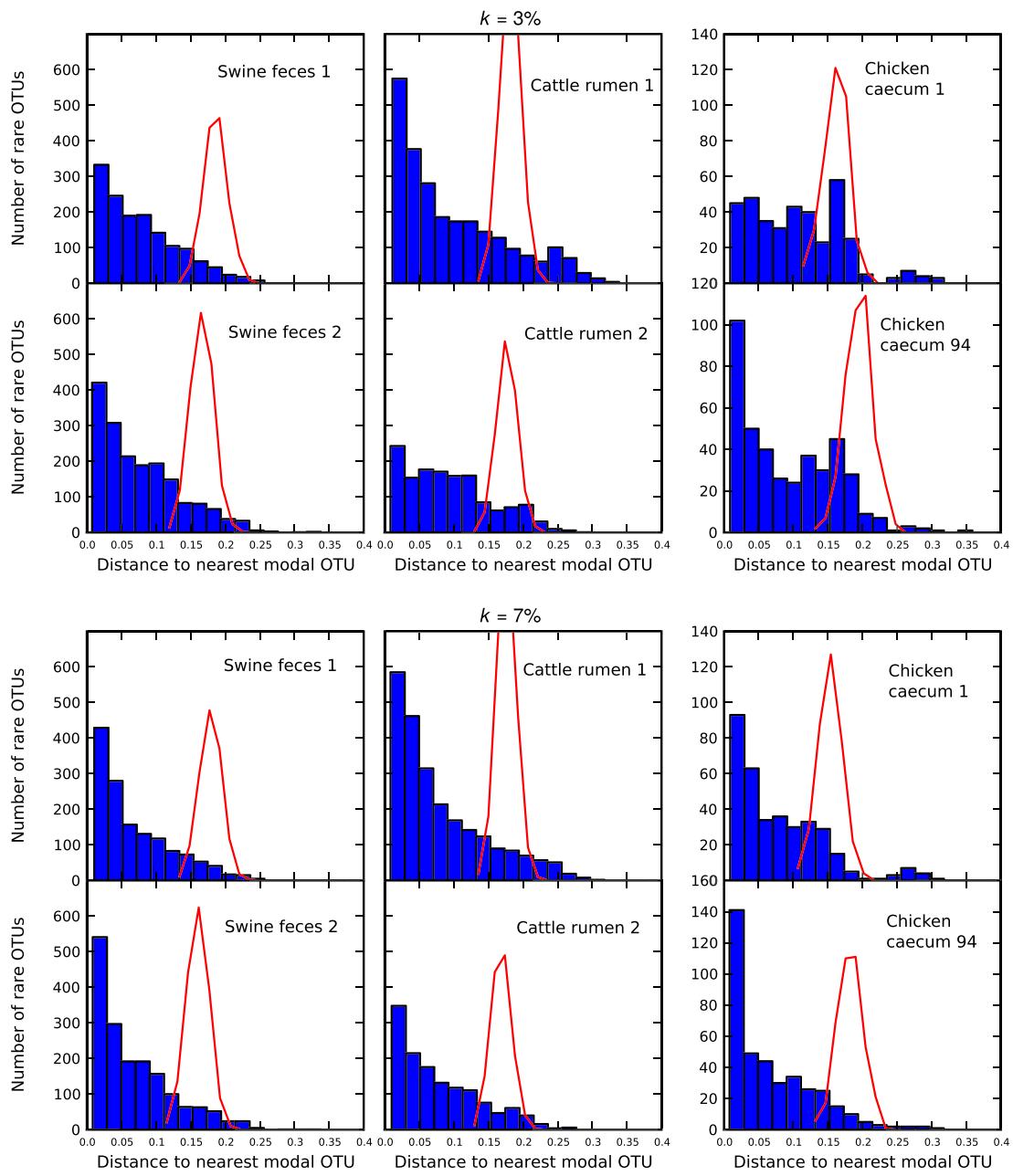
**Fig. S8.** Histogram of distances of rare OTUs to the nearest modal OTU for each of the six GI microbiomes with cutoffs of $k = 3\%$ and $k = 7\%$. Red lines indicate the results of the metric applied to sequences that were randomized while preserving rank abundance and sequence statistics (main text).
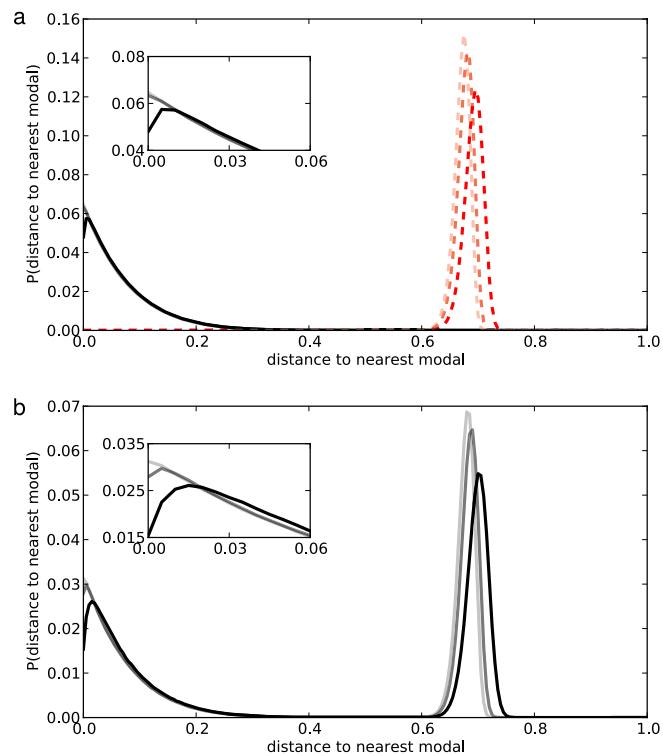
**Fig. S9.** Measurement of the effect of the choice of $k$ on our metric. The darkest lines indicate $k = 2\%$, the medium lines indicate $k = 6\%$, and the lightest lines indicate $k = 10\%$. (A) $\alpha = 0.0$ model (red dashed lines) and $\alpha = 1.0$ model (black solid lines). (B) $\alpha = 0.5$ model (black solid lines).

**Table S1.   Summary statistics of our six datasets**

| | No. of reads | Unique reads | Average length | Aligned width | No. of OTUs at 3% | Simpson diversity | Shannon diversity | Jackknife richness | ACE richness | Chao1 richness |
|---|---|---|---|---|---|---|---|---|---|---|
| Swine feces 1 | 33,283 | 14,122 | 165.0 | 420 | 1,509 | 0.0070 ± 0.0003 | 5.8 ± 0.02 | 2,000 ± 260 | 1,472 ± 55 | 1,540 ± 150 |
| Swine feces 2 | 36,254 | 16,198 | 175.3 | 418 | 1,856 | 0.0068 ± 0.0003 | 5.9 ± 0.02 | 2,300 ± 300 | 1,633 ± 53 | 1,720 ± 150 |
| Cattle rumen 1 | 3,1201 | 18,264 | 180.7 | 471 | 2,580 | 0.0044 ± 0.0002 | 6.3 ± 0.02 | 3,300 ± 260 | 3,070 ± 88 | 2,640 ± 190 |
| Cattle rumen 2 | 19,642 | 10,074 | 183.6 | 385 | 1,509 | 0.0110 ± 0.0006 | 5.9 ± 0.03 | 2,070 ± 110 | 1,818 ± 62 | 1,830 ± 130 |
| Chicken cecum 1 | 17,585 | 2,151 | 136.5 | 310 | 396 | 0.084 ± 0.003 | 4 ± 0.03 | 770 ± 120 | 655 ± 75 | 620 ± 150 |
| Chicken cecum 94 | 21,646 | 2,223 | 138.9 | 332 | 354 | 0.046 ± 0.001 | 3.9 ± 0.02 | 560 ± 90 | 426 ± 57 | 460 ± 100 |