9. Supplementary Material

9.1 *Dynamic unigram simulation experiment*

Next we consider a simulation experiment involving the unigram model. We use the same configurations of D, V, and N as in the LDA and NMF experiments above. The parameter controlling the rate of evolution on the simplex, σ_0^2, is set to 1 throughout. During inference, σ_0^2 is not assumed known, and is instead modeled as an InvGamma $(1, 1)$ across all simulation settings.

In the NMF and LDA experiments, we could visualize fitted posteriors over θ_i and β_j as two-dimensional smoothed scatterplots, since we set $K = 2$. In contrast, this unigram experiment involves a V-dimensional parameter μ_t evolving over time – this precludes any direct analog to Figure 1 or Supplemental Figure 7. Instead, in Figure 10, we visualize the true μ_{tv} against posterior intervals for the one-dimensional $p(\mu_{tv}|x)$, across all configurations of simulation parameters.

Evidently, MCMC provides accurate posterior estimates, while VB and the bootstrap estimates deviate from the true μ_{tv}. First, when the true μ_{tv} is small small, the posterior estimates from VB and the bootstrap are too large. This is not as much a reason for concern as it might appear at first, however, as the parameters μ_t are passed through a softmax before being mapped to probabilities. After this transformation, the difference between, say, $\mu_{tv} = 10^{-2}$ and $\mu_{tv} = 10^{-5}$ is relatively unimportant. Second, posterior estimates for μ_{tv} for large positive values of the parameter tend to be biased downards, to a degree that can't be simply explained as an effect of the prior, as such a strong bias is not present among the MCMC-sampled posteriors.

To simplify the comparison between methods, we can also display the unigram analog of Figure 2 – this is given in Supplementary Figure 11. This display confirms our earlier observation, that posterior MCMC samples seem more reliable than either those from VB or the bootstrap, when using generic probabilistic programming for the Dynamic Unigram model, at least in problems with the dimensions we have considered.

9.2 *Posterior Predictive Checks*

Model assessment is important for qualifying interpretations, and can further guide refinements in subsequent analyses. Indeed, part of the appeal of probabilistic modeling is the ease with which models can be adapted to better describe the data of interest. We briefly review model assessment via posterior predictive checks, as they are applied in Section 4.3. In this approach, some statistics $T_k(x)$ of the data are defined which, in some sense, "characterize" the data. If the data x^* simulated from the fitted model have statistics $T_k(x^*)$ with values similar to those in the observed data $T_k(x)$, then we have evidence that the proposed model approximates the data well, at least in the sense defined by T_k.

More precisely, simulate data $x_1^*, \ldots, x_S^*$ from the posterior predictive probability distribution $p(x^*|x) \approx \int p(x^*|\theta)\,\hat{p}(\theta|x)\,d\theta$, where x is the original data and $\hat{p}(\theta|x)$ is an estimate of the posterior. For each of these simulated data sets, the characterizing statistics $T_k(x_s)$ are computed. Graphically comparing the $T_k(x)$ calculated on the true data with the histogram of model-fit simulated $T_k(x_s^*)$ suggests ways in which the posited model fits – the case where the observed $T_k(x)$ lie in the bulk of the $T_k(x_s^*)$ – or fails to fit – the case where $T_k(x)$ lie far from the bulk of $T_k(x_s^*)$ – the data well.

For example, it is common to set $T_k(x) = \overline{x}_d$ or $\frac{1}{n}\sum_i (x_{id} - \overline{x}_d)^2$ to see whether simulated samples approximately match the moments of the d^{th} dimension in the observed data. Alternatively, histograms of raw data or raw data subsetted to certain groups can guide evaluation. This corresponds to setting a multidimensional $T_k(x)$. For example, $T_k(x_s^*) = (n_{s1}, \ldots, n_{sB})$ could count the number of observations in the s^{th} simulated data set falling into histogram bins $b = 1, \ldots, B$.

9.3 *Comparison with principal coordinates*

There is value in comparing this text-modeling based approach to the data from (Dethlefsen and Relman, 2011) with the analysis carried out in the original publication, based on principal coordinates analysis (PCoA) using a UniFrac distance. Both approaches accentuate the dramatic change in microbiome composition immediately following the administration of antibiotics, and both suggest the overall resilience of the microbiome, in the sense that the community returns to the preantibiotic state after several days. Further, both distinguish subject F as only making an incomplete recovery after the first antibiotic treatment.

On the other hand, some findings are more easily accessible when using the probabilistic approach. For example, the availability of topics with mixed memberships strikes a balance between the continuous gradient representation of principal coordinate analysis of (Dethlefsen and Relman, 2011) and the discrete clusters provided by standard clustering techniques, which appear elsewhere in the microbiome literature. This simplifies the taxonomic characterization of different communities – with principal coordinates, it would be necessary to find species correlated with the principal coordinate axes, and indeed no visual representations of taxa ever appear in the figures of Dethlefsen and Relman (2011). Similarly, by studying individual topics, we find there is more variation within taxonomic families than is ever discussed in the original refernece. Further, through posterior predictive checks, we can perform model assessment in a way that is not so straightforwards with PCoA, and more broadly speaking, we by adopting probabilistic methods, we are able to describe the uncertainty associated with mixed membership and topic estimate.

That said, PCoA enjoys certain advantages over the probabilistic approaches that have been the focus of this work. Perhaps most importantly, PCoA can be run in seconds, which makes it much more useful for interactive analysis. Second, by using the UniFrac distance, PCoA is able to account for the phylogenetic relatedness between taxa, encouraging phylogenetically similar taxa to play similar roles in the resulting ordination. Finally, the visual representation of samples

provided by PCoA – simply a two-dimensional scatter of the samples – is more easily digestible than the simultaneous display of all parameters in probability models.

Broadly speaking, it seems valuable to have both types of tools available for practical scientific work. Ordination techniques like PCoA are useful for describing the relationship between sets of samples, in a way that requires little effort in either estimation or display. However, for more richly structured summaries, which are amenable to uncertainty quantification and model evaluation, probabilistic methods are ideal. We imagine a workflow in which researchers quickly develop a sense of their data using ordination, and then refine and critique their analysis using latent variables models.

9.4 *Supplementary Figures*

[*Received XXXXXX, 2017; revised XXXXXX accepted for publication XXXXXX*]

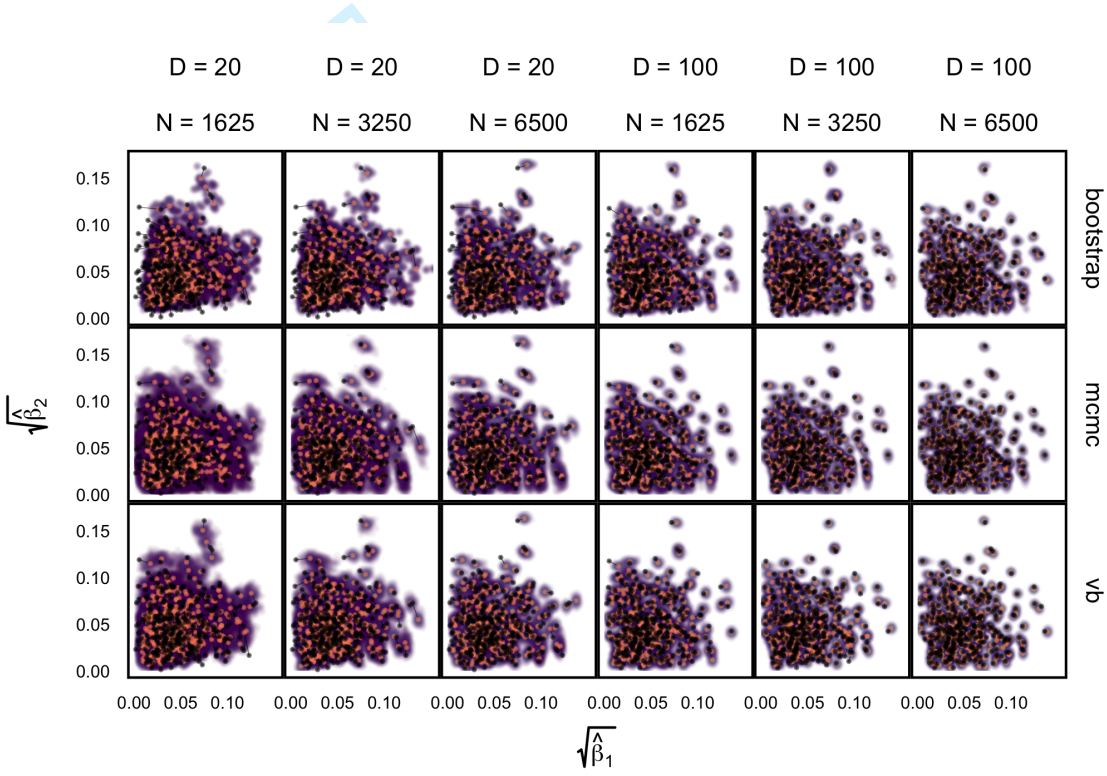

Fig. 6. The analog of Figure 1 in the case that $V = 325$.

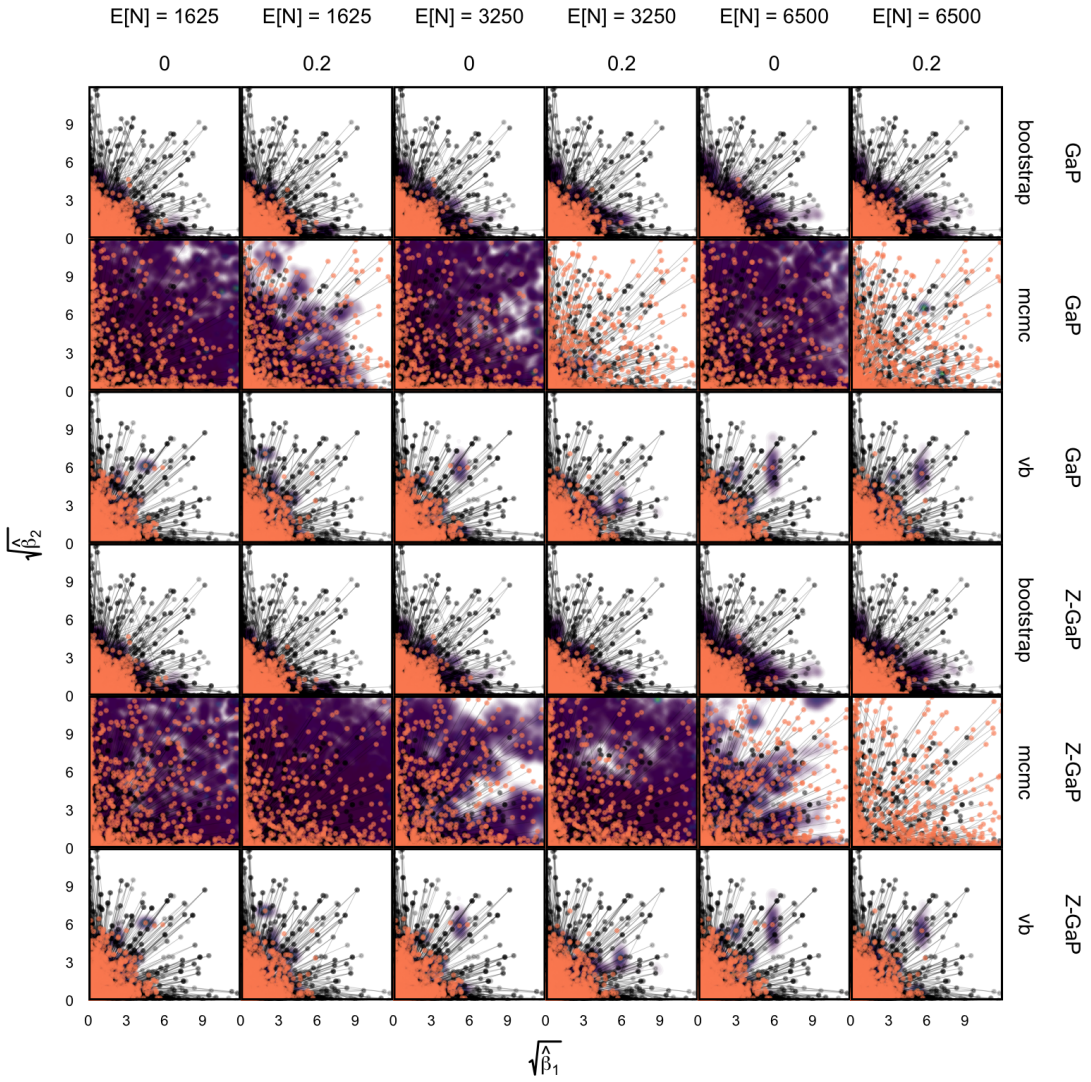

Fig. 7. The results of the NMF experiment with $D = 20$ Within each panel, we display the true value of $\sqrt{\beta_v}$ as black points, while linked orange points give associated posterior medians. Note that the axes are truncated, and for some panels, the posterior medians all lie outside the visible box. Across columns, we vary and $\mathbb{E}[N]$. Along rows, we vary the assumed model, the inference procedure, and the true p_0 – these are the three columns of row labels, read from outside in.

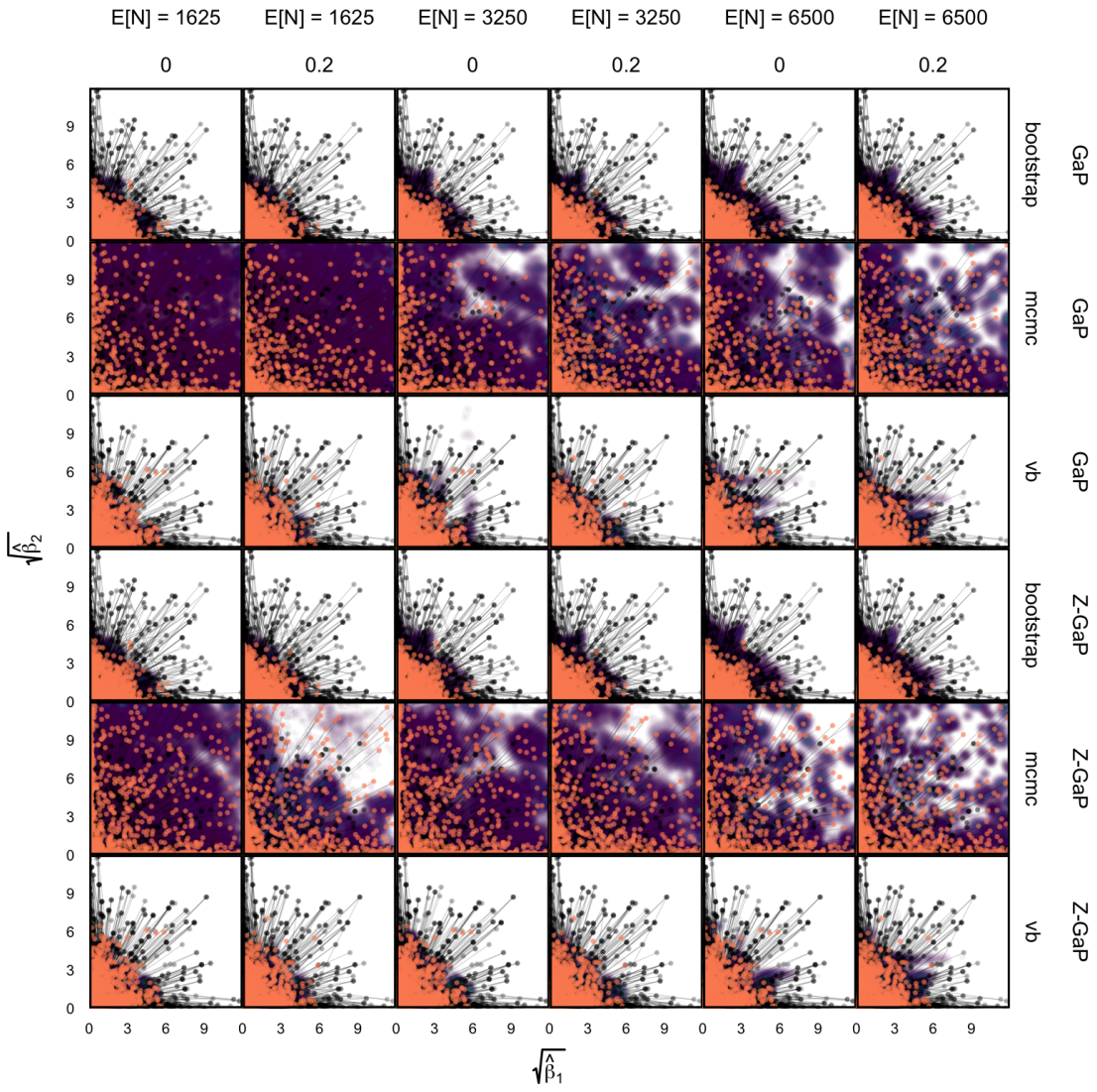

Fig. 8. The analog of Figure 7 when $D = 100$.

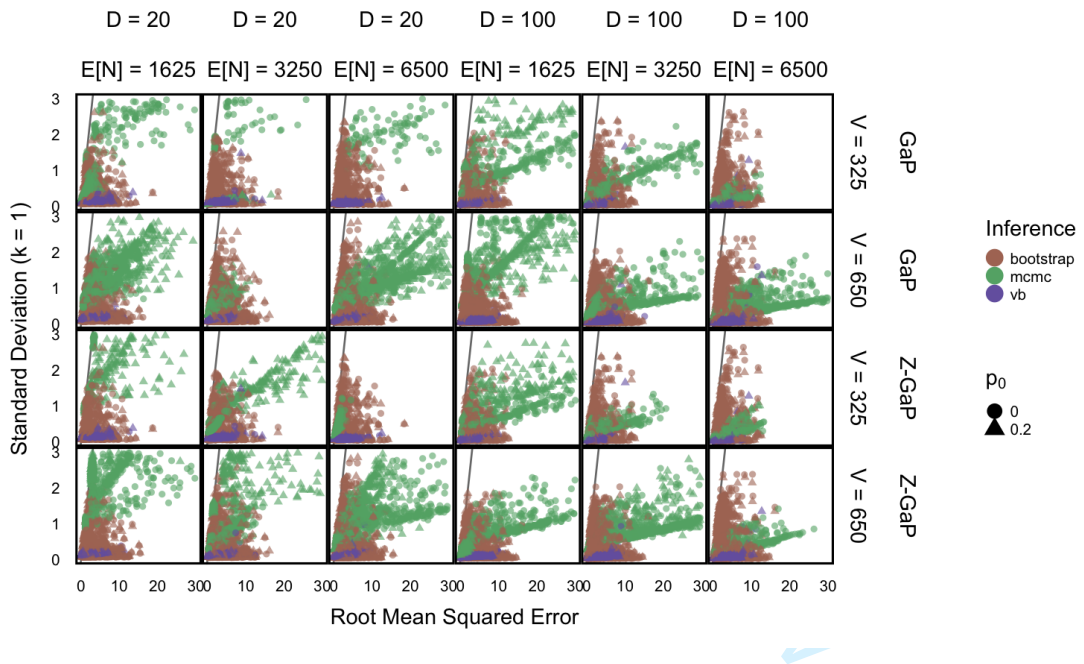

Fig. 9. A version of Figure 2 for the NMF simulation experiment. The figures are read similarly, except there are a larger number of experimental configurations – rows now distinguish the assumed model and shapes represent the true value of p_0. Further, while the first row of column labels still gives D, the second row gives $\mathbb{E}[N]$ instead of N. Note that we also now truncate the x and y axes, and not all points are visible. For example, most MCMC samples in the last panel in the second row have errors and SDs larger than what is displayed, and so are missing.

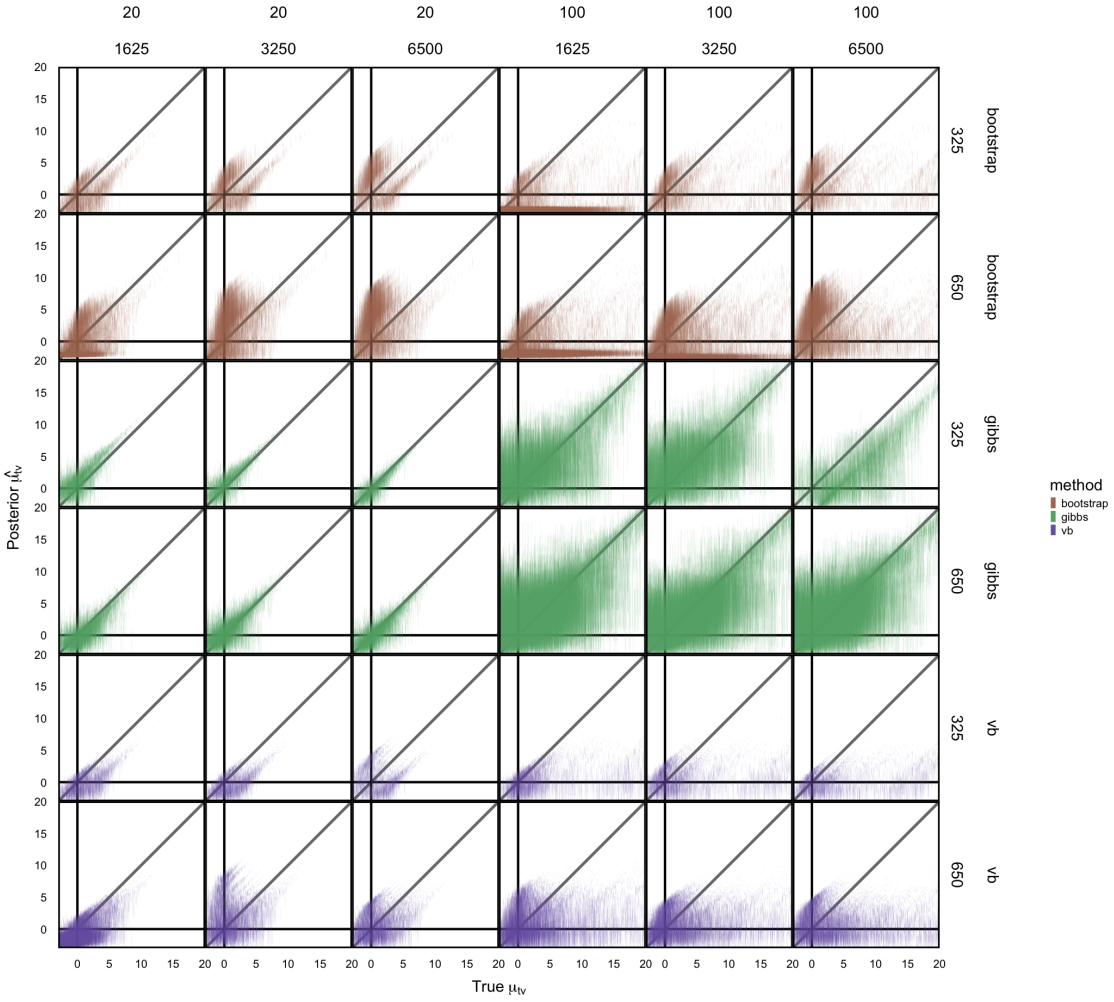

Fig. 10. A comparison of the posterior $p(\mu_{tv}|x)$ to the known underlying μ_{tv}, in the unigram simulation experiment. The x-axis for each interval corresponds to the true μ_{tv} for one species at one timepoint, while the vertical intervals cover the 25% to 75% quantiles of samples from the posterior $p(\mu_{tv}|x)$. Different panels distinguish between configurations of D, N, V, and posterior sampling schemes. Posteriors from MCMC sampling seem to correctly recover the true underlying μ_{tv}, while discrepancies arise for both VB and the bootstrap.

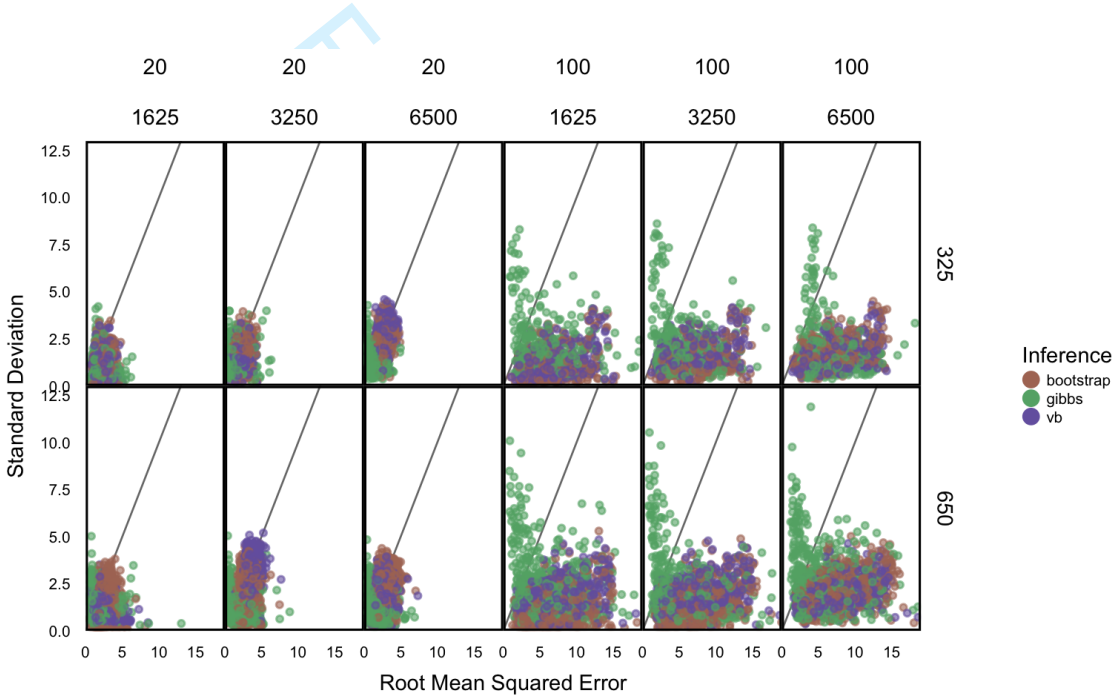

Fig. 11. A simplification of Figure 10, displaying RMSE when using posterior medians to estimate simulation μ_{tv}s (x-axis) and the standard deviations of posterior marginals (y-axis), across experimental configurations. Generally, MCMC sampled posteriors seem to be the most reliable, across simulation configurations.

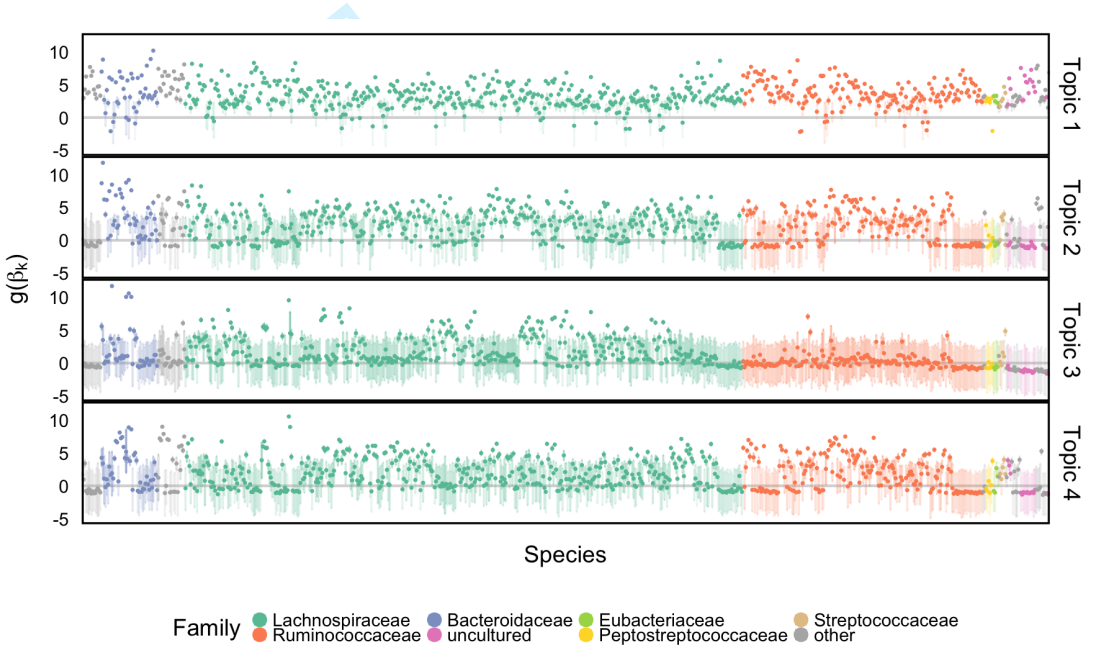

Fig. 12. Each credible interval describes an approximate posterior for one β_{vk}. Coupled with Figure 3, this guides the interpretation of which bacterial taxa are more or less prevalent during antibiotic treatments. Each row of panels corresponds to one of the four topics, the x-axis indexes species, sorted according to phylogenetic relatedness, and the y-axis give transformed values of the species probability under that topic. Only the 750 most abundant species are shown. Note the disappearance of otherwise abundant species within Topics 2, 4, and to some extent, 1.

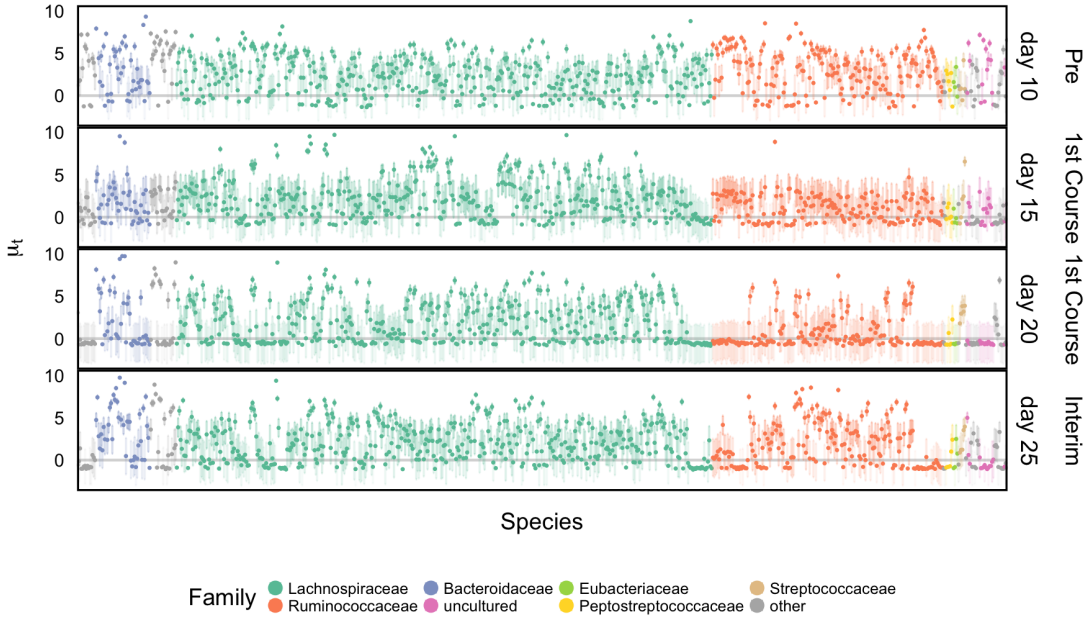

Fig. 13. Each posterior credible interval refers to one μ_{vt}. The rows are a subset of times t around the first antibiotic time course. The first row corresponds to a timepoint from before the treatment, the middle two from during the antibiotics time course, and the bottom from after the time course was stopped. Otherwise, this display is read in the same way as Supplementary Figure 12. This view provides one way of smoothing abundance time series, to see how different species respond to antibiotic treatment.

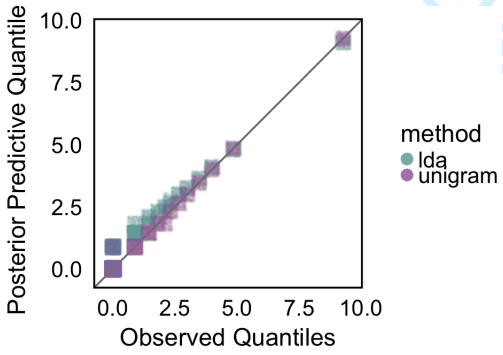

Fig. 14. As a posterior check, we compare the observed with simulated data quantiles, using a qq-plot. To reduce overlap, we have introduced a uniform $[0, 0.2]$ jitter on both axes. Further, the points are semi-transparent – this makes it easy to see that most quantiles map to 0, which is expected, considering the sparsity of the data. From this view, we see that the LDA model tends to underestimate the overall number of zeros in the data, while the Dynamic Unigram model matches the observed quantiles almost exactly.

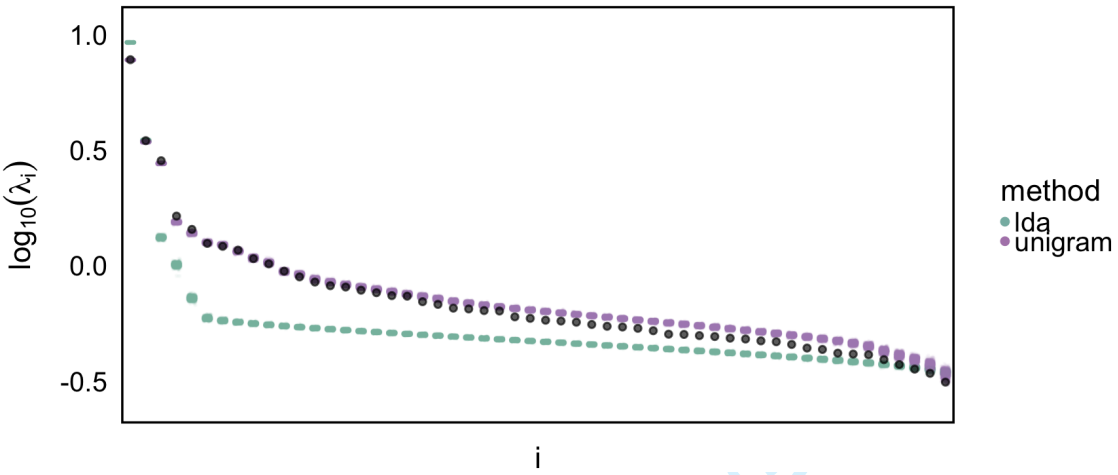

Fig. 15. As a posterior predictive check, we compute eigenvalues of data simulated from the fitted LDA model. The clouds of points summarize the posterior predictive distribution, while the black circles represent observed data eigenvalues. Note that the y-axis are logged eigenvalues. Evidently, the four-topic model effectively creates a rank-four approximation of the original data.

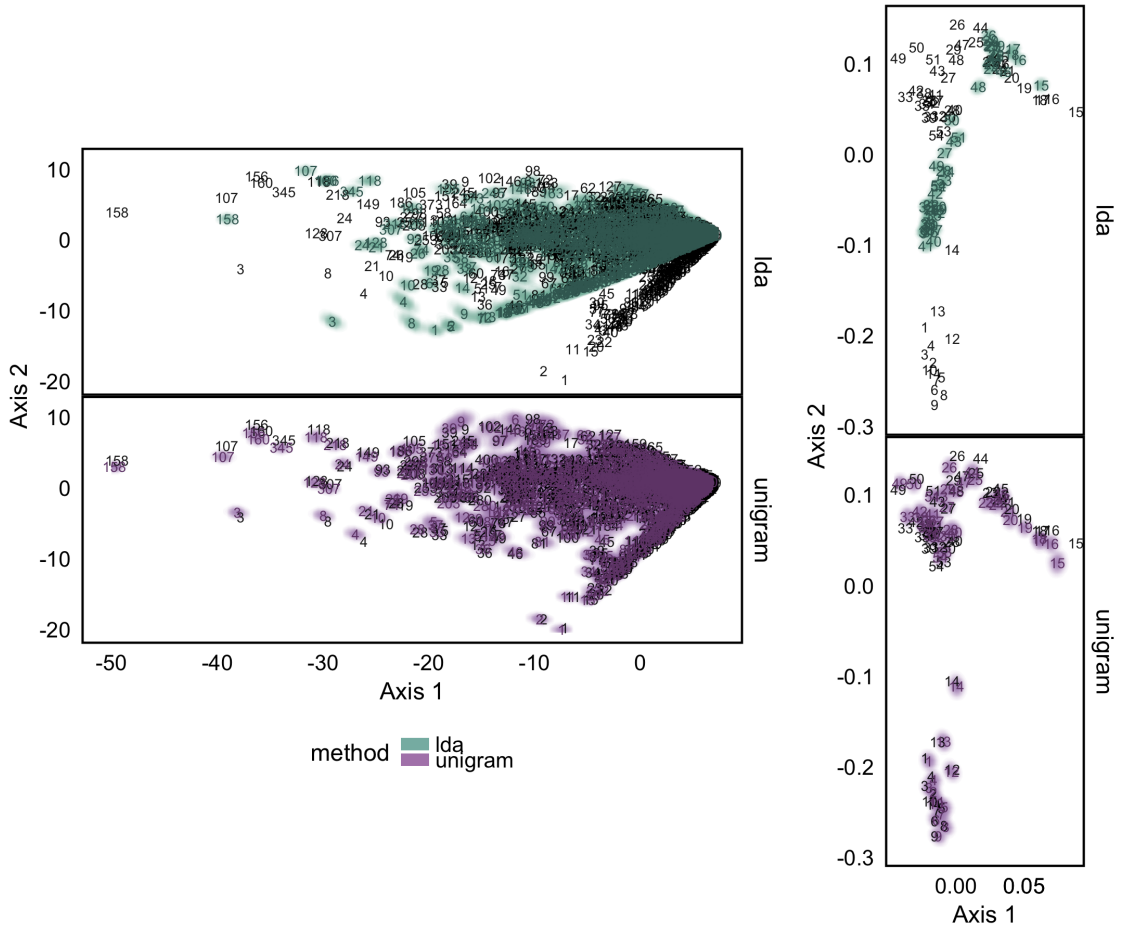

Fig. 16. The eigenvalues displayed in Figure 15 correspond to PCA results computed on posterior predictive samples, which are aligned and overlaid here. The left pair of panels give scores for each species, while the right pair provide loadings for each timepoint. The individual posterior samples have been smoothed into contours, while the posterior medians are displayed as shaded text. The observed data PCA results, after alignment with posterior samples, are displayed as black text.

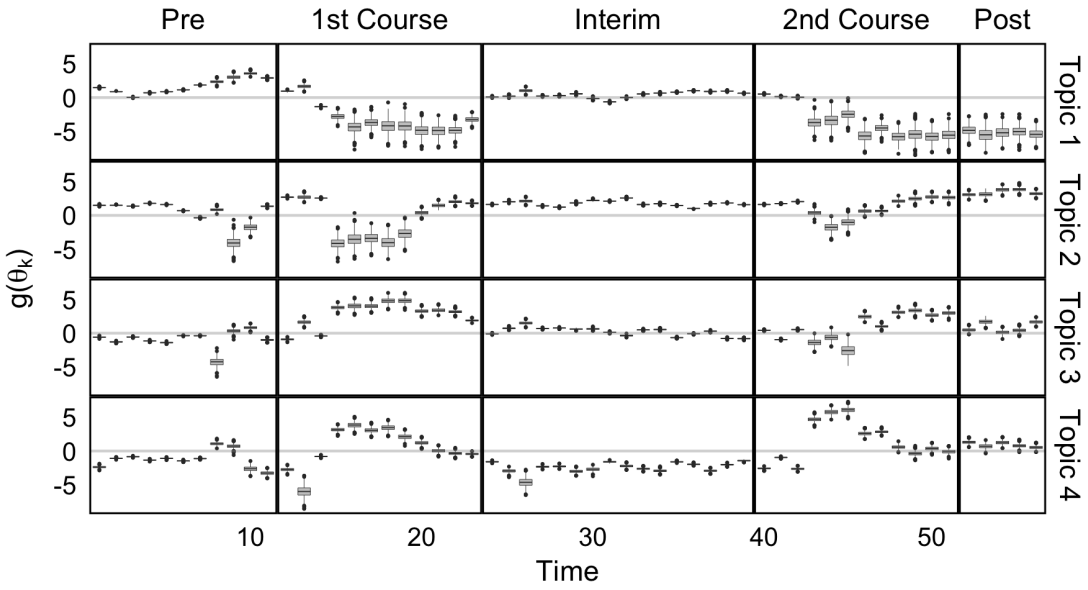

Fig. 17. The analog of Figure 3 for Subject D.

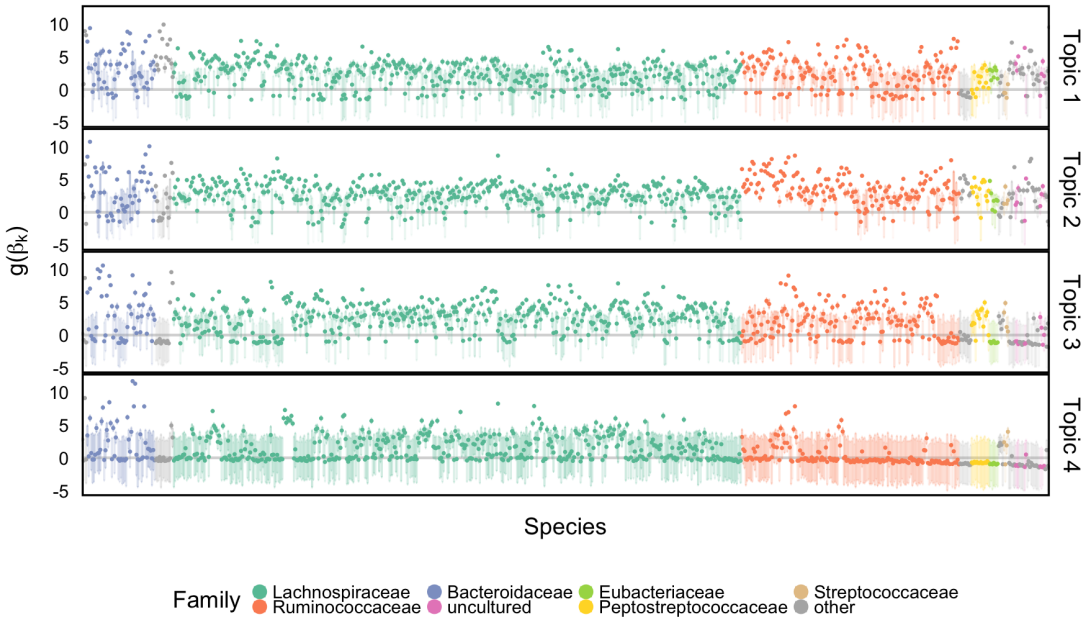

Fig. 18. The analog of Supplementary Figure 12 for Subject D.

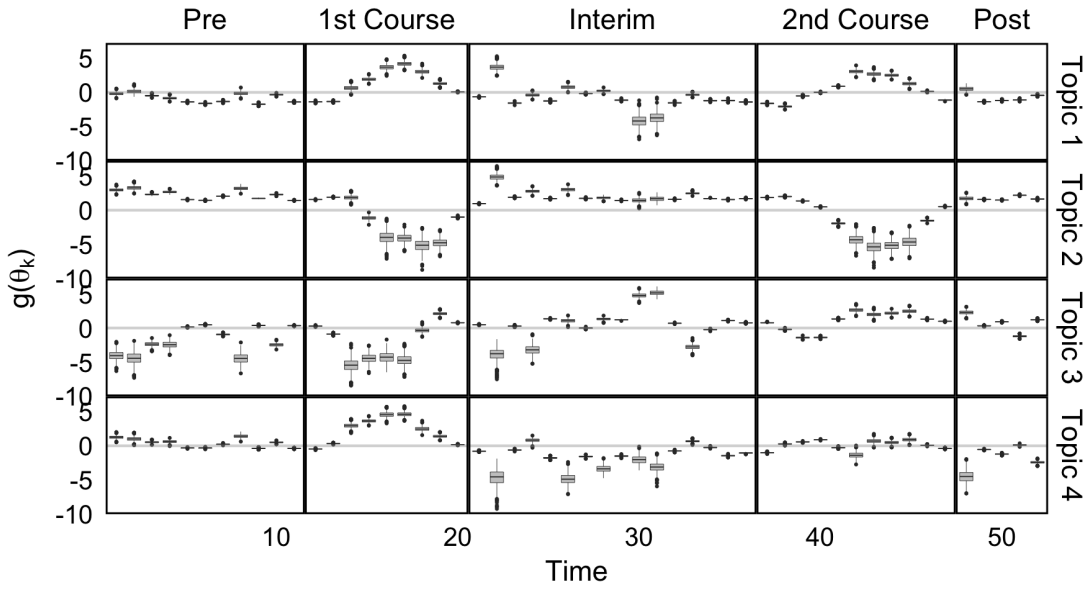

Fig. 19. The analog of Figure 3 for Subject E.

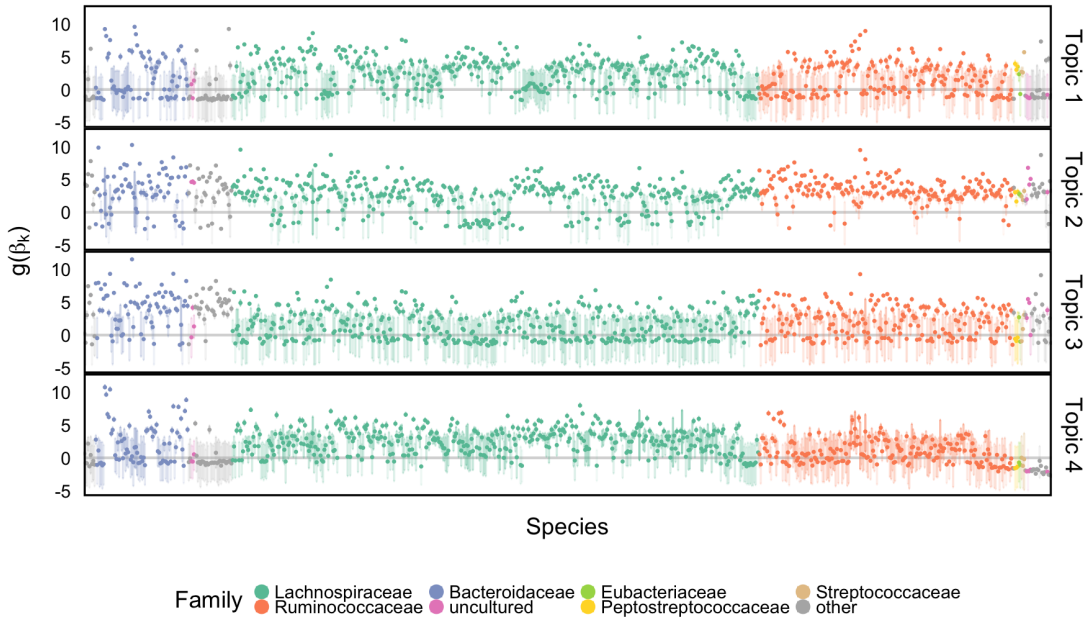

Fig. 20. The analog of Supplementary Figure 12 for Subject E.

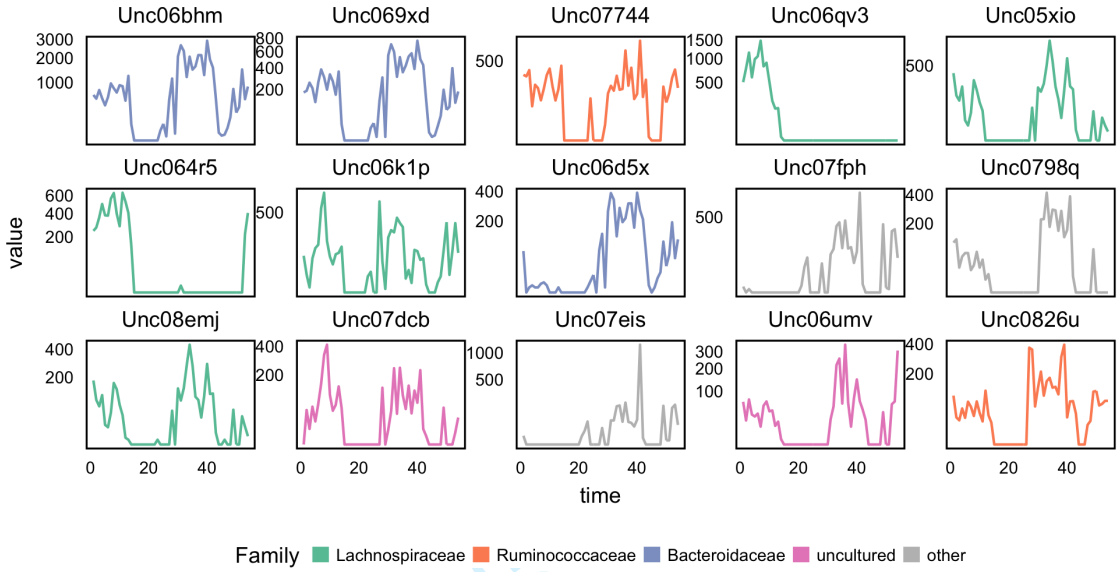

Fig. 21. Rather than displaying all representative species together, as in Figure 4, we can sort species according to how representative they are of an individual topic. Here, the 15 species most strongly associated with Topic 1 are given. The panels are to be read from left to right and from top to bottom, to go in decreasing value of association. Note that the y-axis is on a square root scale.

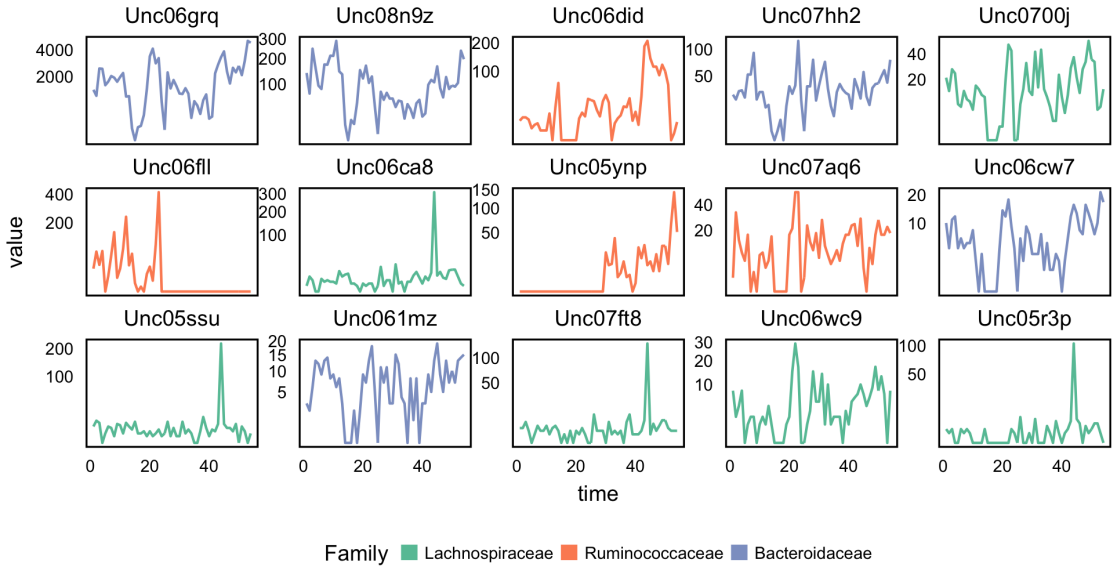

Fig. 22. The analog of Figure 21 for Topic 2..

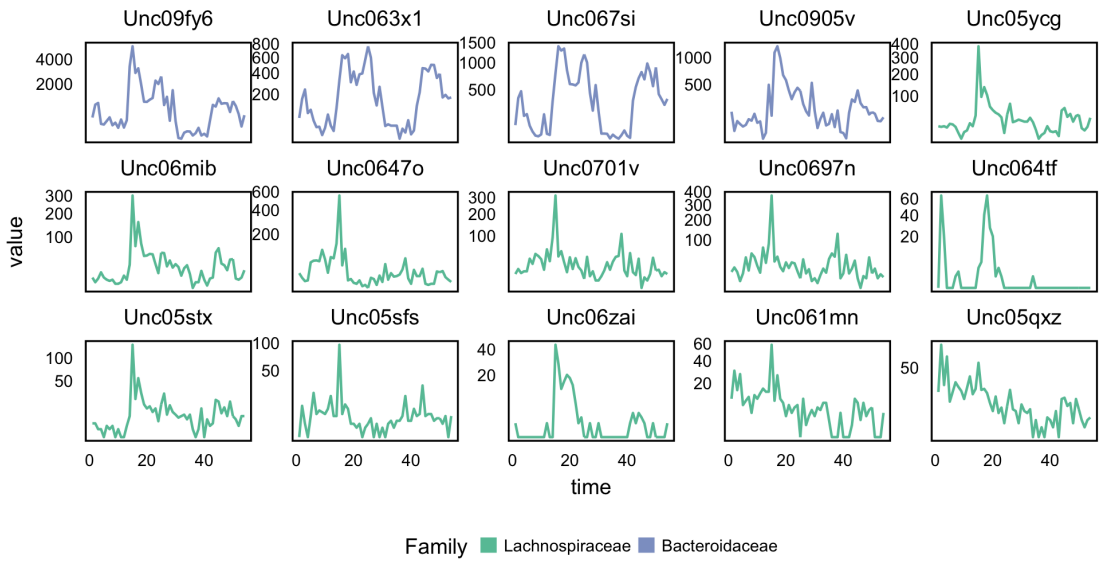

Fig. 23. The analog of Figure 21 for Topic 3.

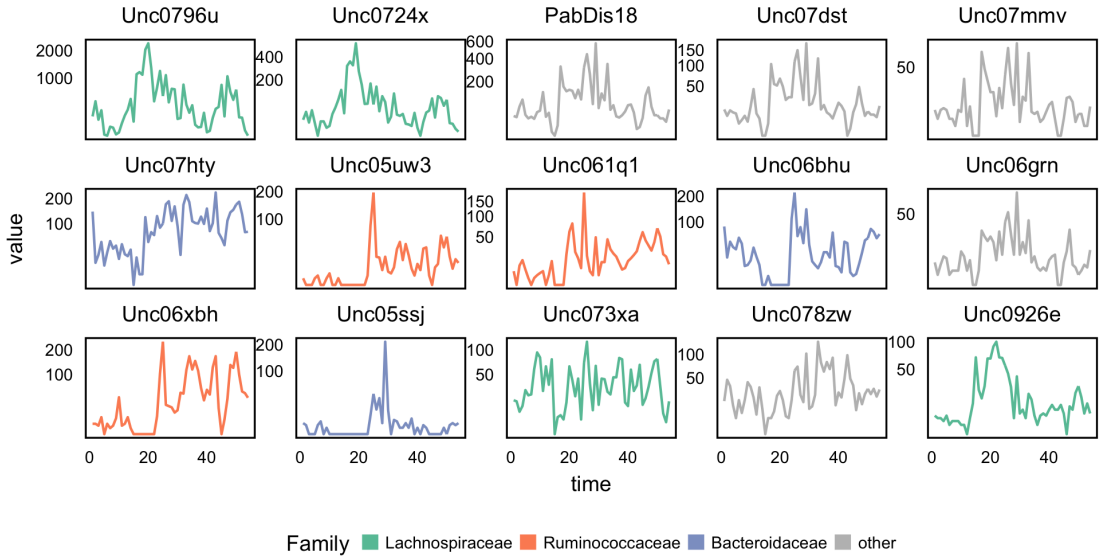

Fig. 24. The analog of Figure 21 for Topic 4.

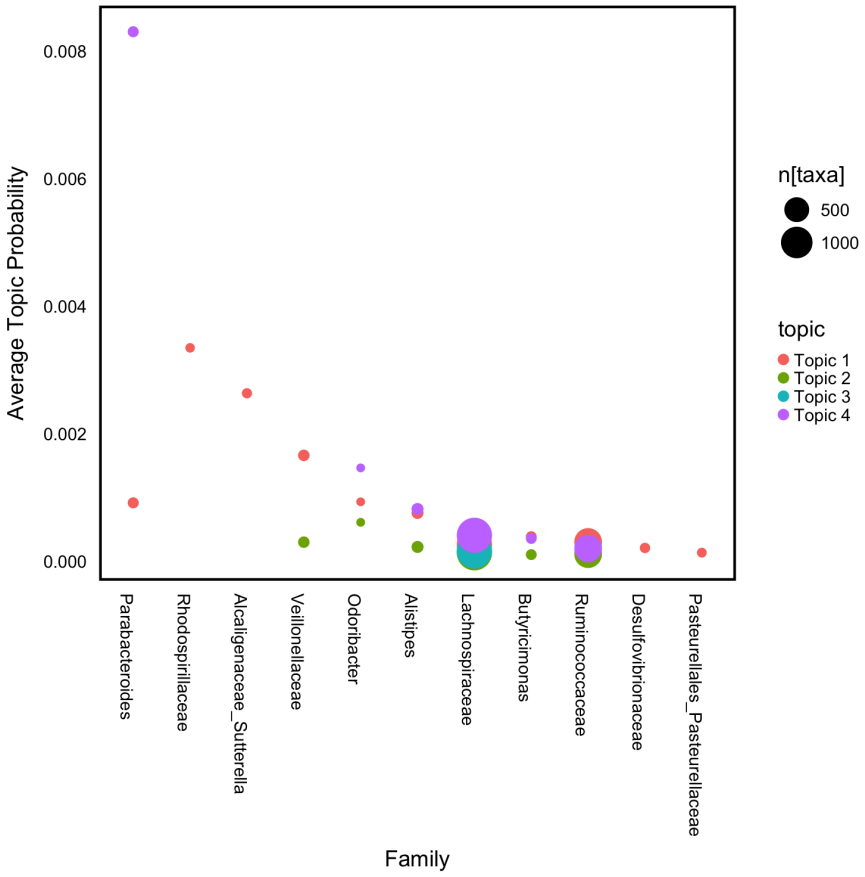

Fig. 25. We can search for entire taxonomic families that seem associated with individual topics. Here we calculate the average of the topic representativeness statistic $\beta_{kv} - \sum_{k' \neq k} \beta_{k'v}$ across all species v within each Family. Only those families that are most associated with a topic are displayed here. The sizes of circles represents the number of species within the family, which can be used to gauge the variability of the estimate. Compare this view with Supplementary Figure 26.

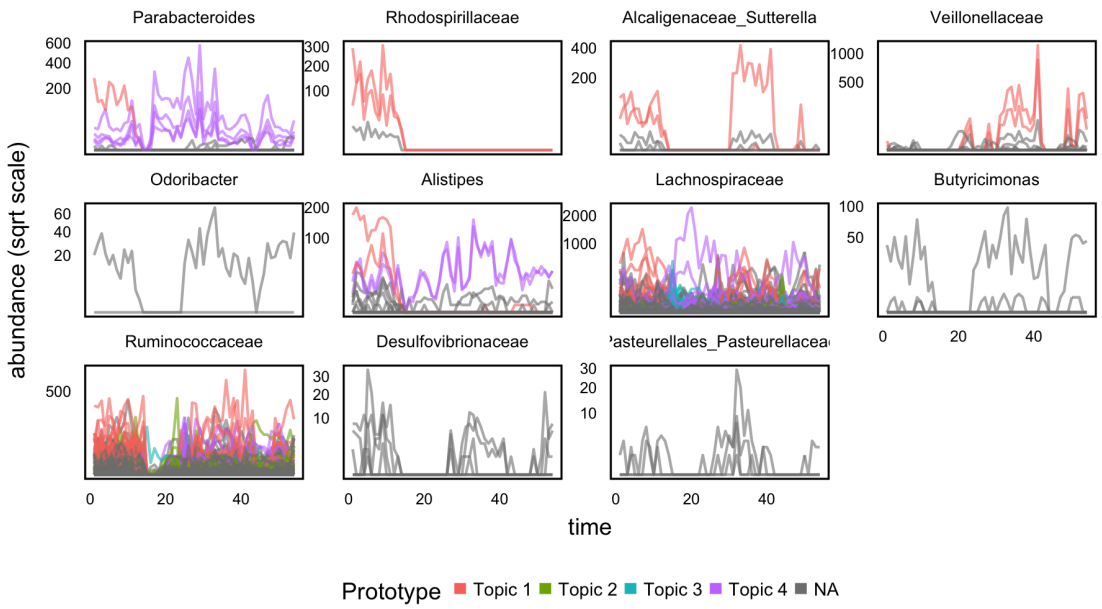

Fig. 26. All species within the families screened out from Figure 25 are displayed here. Species among the representatives displayed in Figure 6 are colored according to the topics of they are prototypical. Grey series still contribute to the average family-topic association measure, but were not among the 50 prototypes for each topic. This view suggests that the Rhodospillaceae, Alcaligenaceae and Parabacteroides may have a large fractions of representatives from Topics 3, 1, and 2, respectively. Note that as before, abundances are plotted on a square root scale.