Notes on RAD MS

**Introduction**

Species abundances in a community must be derived from a few basic processes: (1) species’ interactions with their environment, (2) species’ interactions with other organisms, and (3) migration of individuals into and out of the community. It has been a longstanding question in ecology, however, of how best to characterize the distribution of species abundances in the community.

Something about SADs and RADs…

Most studies of species abundance distributions have focused on the problem of which model best fits observed patterns in abundance. Very rarely are the parameters of the best fitting model explored. Yet we have reached the inescapable conclusion that there is most proposed models fit observed abundance distributions reasonably well. A better approach may be to instead determine what leads to changes in the various properties of the abundance distribution (Matthews et al. 2017).

Hughes 1986 demonstrated the usefulness of a dynamics-based approach, finding that a large proportion of sampled communities display abundance curves that were qualitatively similar to those produced by his relatively simple dynamic model. More recently Powell and McKane (2008) used a dynamic food web model to simulate communities in order to generate an *in silico* dataset that they then used to compare species abundance distributions. They found that certain model features could be used to determine which distribution best fit to their data. By using simulations, the authors were able to resolve questions that would be exceedingly difficult to accomplish in the field.

From Hughes 1986:

*The dynamics model predicts the species-abundance pattern of most samples with greater accuracy and provides an explanation of species abundance based on recognized and testable ecological principles.*

Gaining a complete understanding of the processes underlying the population dynamics of a community is incredibly difficult. By generating community data through simulation, however, we can assess how different underlying processes generate realistic distributions. Using a Lotka-Volterra (LV) model allows us to address how species interactions with their environment and each other may affect the distribution of abundances in the community (for our purposes we ignore migration, although it could conceivably be incorporated into the model). In this model species interactions with each other are explicit, and their interaction with the environment is implied through the implementation of growth rates, carrying capacity, and “self” interactions.

An important distinction between this study of abundance distributions and those of others is that we explicitly look at communities of all interaction types, while most studies of abundance distributions have focused on horizontal communities only (e.g., all plants in a given area). The restriction to horizontal communities limits the potential influence of species interactions, as these communities typically expected to only exhibit competitive or facilitative interactions.

We seek to answer three main questions about the rank-abundance distribution (RAD) using simulations based on the LV model.

1. What is the effect of interactions with the environment on species rank abundance distributions?
2. How does variation in species interactions drive variation in the shape of their abundance distribution?
3. Can the LV model generate communities with species RAD that closely mirror those of observed communities?

We will answer the first question in two ways. First, we will use data compiled by the American Gut Project to determine how gut-environmental variables based on health, diet, and exercise affect the shape of the abundance distribution. In addition, we will examine how implicit environmental variables included in simulations alter the distribution. To address the second question, we will use linear regression to determine how the number and strength of different species interaction distributions in the community affect the fitted RAD. We also include species’ growth rates, self-interactions, and total number of species in our regression. The last question will be addressed using classification and regression tree analysis, paired with random forest models. These methods will highlight the importance of the different predictor variables in generating communities that lie either in or out of the parameter range of naturally occurring communities.

The Zipf model has been shown to fit reasonably well to many species abundance distributions (but see Baldridge for otherwise), and may easily be used in the rank-abundance form using the `sads` R package. This RAD model has a single parameter *s* that describes the shape of the distribution, where lower values indicate more evenly distributed abundances.

**Methods**

*Data*

We compiled human gut microbiome data from various sources, including the Human Microbiome Project (HMP) and the American Gut Project (AG). Data from the HMP were downloaded as a final OTU table output following processing in QIIME from the project website (<http://hmpdacc.org/HMQCP/>). We used the OTUs picked from the sequenced V1-3 region. This dataset includes samples from 187 individuals covering 8048 OTUs in total. The latest AG data were downloaded from the project’s ftp site (<ftp://ftp.microbio.me/AmericanGut/latest/11-packaged/fecal/100nt/all_participants/all_samples/10k/>). We used processed data from all individuals and all samples that had been rarefied to 10K reads. These data include 6108 individual samples covering 24443 OTUs.

In addition to data from the HMP and AG, which represent snapshots of the gut community in time, we compiled time-series relative abundances from four sampled individuals (time-series lengths range from ~250 days to 360).

*Simulation*

To create *in silico* communities for simulations, we generated undirected random networks with the Erdos-Renyi model. In this model, all species have an equal probability *C* of interacting with any other species. The probability *C*, also termed the connectance of the network, was drawn from a uniform distribution between 0.1 and 0.8. There is very little information about the true connectance in microbial communities, so we chose values that reflect information from two sources. In macrobiological networks (e.g., food webs) connectance tends to be low, between 0.1 and 0.3 (Dunne et al. 2002). However, recent work using metabolic reconstructions of human gut microbiota to conduct pairwise growth experiments *in silico* suggests that most microbes interact with one another, with 80% of pairwise simulations resulting in an interaction (Magnúsdóttir et al. 2017).

Interactions in each community were defined by assigning a +, -, or 0 to each link. For each network these signs were given different probabilities drawn from a uniform distribution. Because the original network was undirected, if species A interacts with B then B interacts with A. By giving either a sign or zero to each of these links were able to include all five major interaction types (parasitism, mutualism, competition, amensalism, and commensalism) into the network. The signed interactions were then given strengths drawn from the absolute value of a normal distribution with mean of 0 and standard deviation of either 1 or 0.5. Intraspecific interactions were drawn independently from a uniform distribution between -2 and 0.

We assumed that population dynamics followed the generalized Lotka-Volterra equations,

where *ri* is the growth rate, *Ki* is the carrying capacity, and *aij* is the effect of species *j* on species *i*. Species growth rates were all set to be positive, and drawn from a uniform distribution between 0 and either 0.2, 0.5, or 1. The carrying capacity was set to be equal for all species. Its value was determined by dividing a community-level carrying capacity by the number of extant species. Our simulations used carrying capacities from 20 to 200. As an example, if the community-level capacity was set to 20, and there were 100 species extant in the community, then *Ki* would be 0.2.

All simulations were completed using R version 3.4.3. Projection of population dynamics through time were accomplished using numerical integration with the ode function in the deSolve package. We simulated dynamics for XXXXX communities.

*Rank Abundance*

Why use rank abundance?

We chose to use the Zipf model to characterize the rank abundance distribution. The model was fit to both real and simulated data using maximum likelihood methods. To facilitate comparison across studies, each sample was rarefied to 2000 individuals (i.e., reads).

*Statistics*

**Results**

Rank Abundance Distribution fits

* Real data
* Simulated data

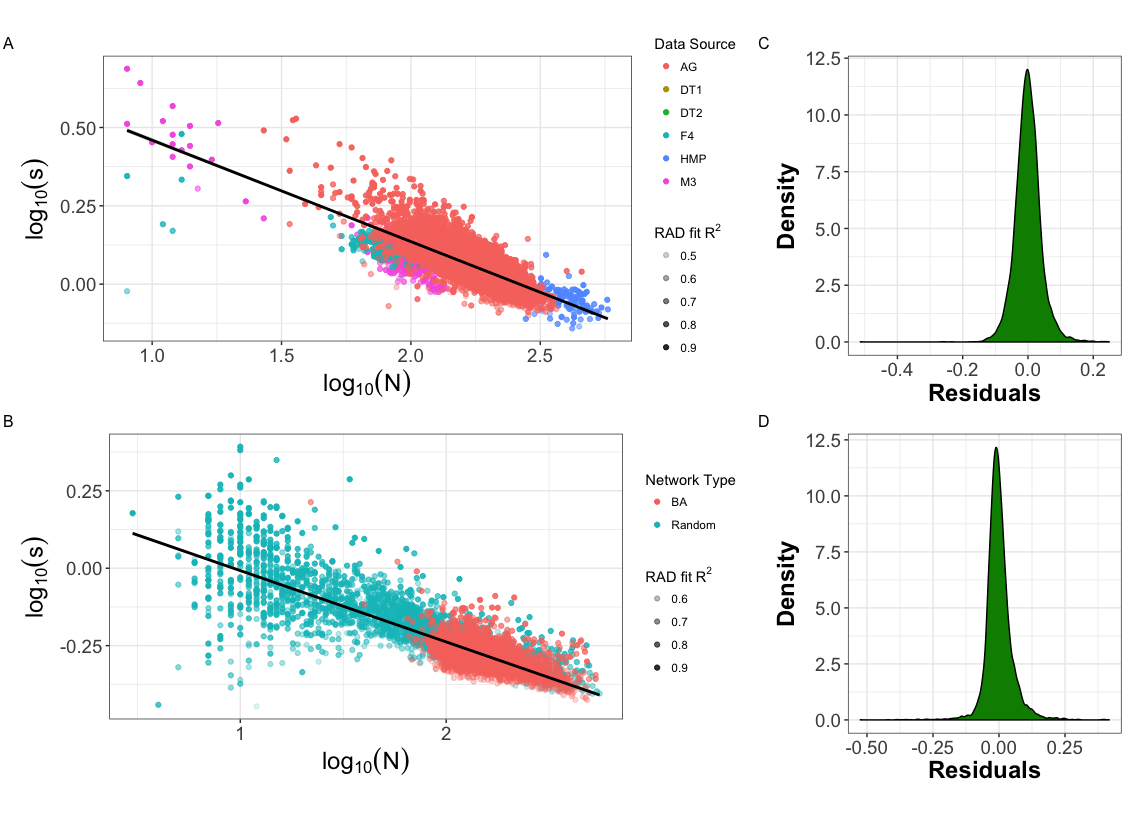
How well do species interactions explain variation in s?

* Our regression explains a significant amount of the variation in fitted *s* values to simulated data.

Does simulated RAD match up to real RAD fits?

Are there alternative explanations for variation in s? How does environment (e.g., AG metadata) affect s?

* Interesting relationships between s and lifestyle variables
  + Lower average s in healthy individuals
  + Higher average number of species in healthy individuals
* Effect of diet on s or max/total abundance



**Figure 1:** Plot of the fitted Zipf parameter *s* against the number of species in the community (*N*). (A) Log-log plot of s against N for gut microbial communities from the American Gut Project (AG), Human Microbiome Project (HMP), and time series data from Daily Timescales and Caporaso. (B) Log-log plot of s vs N for simulated communities whose interactions networks were either hub-like (Barabasi-Albert model, BA) or random (Erdos-Renyi model). Black lines in A and B represent linear regressions of log10(s) ~ log10(N). Residuals for each regression are plotted in C (from A) and D (from B).

**Discussion**

Zipf fits well, can be substantial variation across and within individuals.

A few reasons why simulated distributions do not match observed distributions are that: (1) observed communities are likely an amalgam of several spatially structured communities throughout the gut because of the sampling method, (2) migration into and out of human gut microbial communities, (3) copy number variation in the 16S RNA gene may be skewing results, (4) current methods do not adequately sample the gut community, or (5) clustering of OTUs does not accurately represent the true species distribution. On the other hand, a number of assumptions inherent in the simulation process may also be to blame. For example, we simulated initially randomly interacting communities while most real communities in macrobiological systems are decidedly non-random. We also do not know the true distribution of interactions or their strengths in real microbial communities, so we may have chosen incorrect distributions for these parameters, or not adequately sampled parameter space.

We have also likely not captured the full extent of the influence the environment may have on the abundance distribution. Without explicitly modeling this interaction (e.g., through consumption of allochthonous resources), we are limited in our ability to determine the effect it has. The species’ growth rate and self-interaction term may not be adequate proxies for their interaction with the environment. No host-microbe interactions, no migration (i.e., transient vs. core species).