Estimation of species richness

Peter Sutovsky

School of Mathematics & Statistics,
Newcastle University, UK

26th May, 2017

Biodiversity and metagenomics

- Biodiversity (”biological diversity,”) – variety and variability of life on Earth.
- Method: metagenomics (environmental genomics, ecogenomics, or community genomics)
    - Study of genetic material recovered directly from environmental samples (Wikipedia)

PROBLEM: METAGENOMIC SURVEYS RECOVER ONLY A SMALL FRACTION OF THE EXTANT DIVERSITY. NONETHELESS, MANY METHODS TREAT THE OBSERVED SAMPLE AS THE POPULATION.

Microbes

- Microbes are required to sustain almost every other form of life and influence
    - Climate, health, and agricultural productivity and the fate of pollutants
    - Sometimes unanticipated modulators
- Laboratory cultures different form real life microbial communities
- How do these microbial communities form?

Microbial diversity

- Diversity – the number of taxa or species richness as well as their relative abundance.
- Operational taxonomic units (OTUs) – based on differences of 16S rDNA sequences
- Taxa – clusters of sequences that differ by at most 3% of sites
- Microbial world - 10^{30}organisms
  - Vast, diverse and largely unexplored
  - Observed through relatively small sample size
- Sample size - Technological and financial limitations
- How big a sample is big enough?

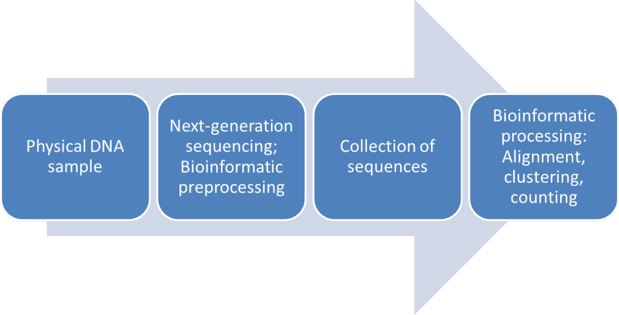
Microbial ecology and metagenomics

- Microbial ecology (or environmental microbiology) is the ecology of microorganisms
  - their relationship with one another and with their environment.
- It concerns the three major domains of life – Eukaryota, Archaea, Bacteria, and viruses
  - fingerprinting of microbial communities or assessing biodiversity
- Microbial ecology and biotechnology provide tools to address environmental and economic challenges
  - e.g. for fingerprinting, assessing biodiversity, and tracking the changes of microbial communities

Species abundance

- Relative species abundance and species richness describe key elements of biodiversity
- Relative species abundance
  - component of biodiversity and
  - refers to how common or rare a species is relative to other species in a defined location or community

Abundance data



Physical DNA sample

Next-generation sequencing; Bioinformatic preprocessing

Collection of sequences

Bioinformatic processing: Alignment, clustering, counting

Cluster sequences at some % “identity,” typically 97%
{clusters} = {OTUs} OTU = “operational taxonomic unit”

Species diversity

- Comprised of:
  - *species richness*: number of species present
  - *heterogeneity of species*
    - relative abundance of each species present in the community

Statistical problem

- Estimate total population diversity – number of species, classes, taxa, OTUs – based on frequency count data
- Data =
  - # of units observed exactly once in sample (singletons);
  - # observed exactly twice (doubletons);
  - # observed exactly three times; $\ldots$.

Frequency count data example

Global Ocean Survey (GOS) data from the upper oceans given by Rusch et al. (2007)

Abundance	No. of species
1	311
2	213
3	61
4	38
5	33
...	...
34	1
36	2
38	1
39	1
...	...
365	1
1163	1

Analysis of species abundance data - whole population

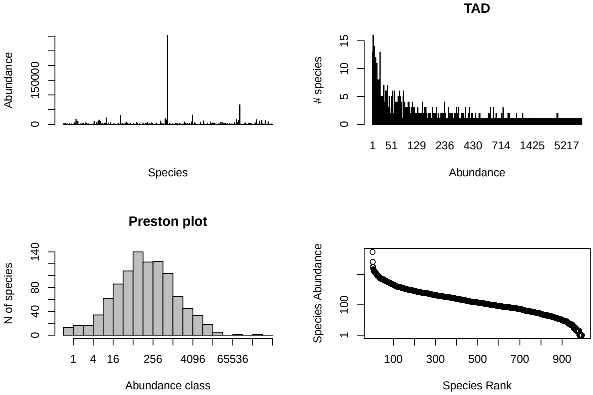


Figure: Abundance of species - whole population

Analysis of species abundance data - 0.02% of community sampled

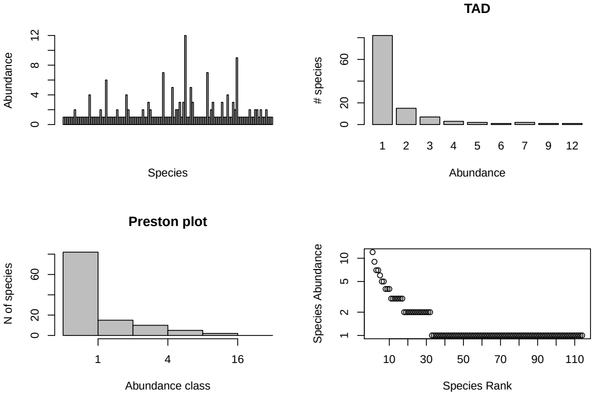


Figure: Abundance of species - 0.02% of community sampled

Taxa abundance – real TARA ocean data

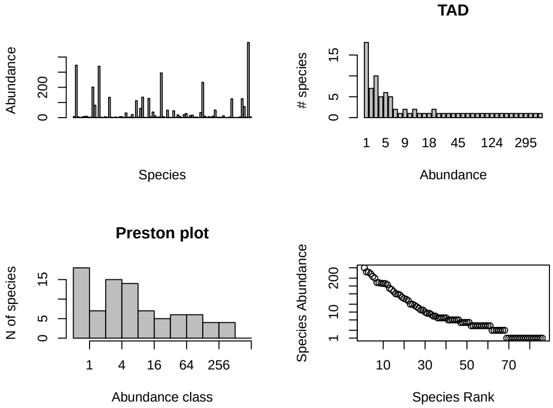


Figure: TAD of TARA ocean data (ERR315852)

Diversity estimator function

i – abundance

- $P_0 = P(n = 0|\boldsymbol{\theta'})$
- f_k – number of taxa with abundance k
- S – total number of taxa in the community
- $D = \sum_{k=1}^{L} f_k$ – number of observed taxa in sample
- $f_0 = S - D$ – number of unobserved taxa in sample

$$\hat{S} = D + \hat{E}[f_0]$$

$$E\left[f_0\right] = \sum_{i=1}^{S} P\left(n_i = 0\right)$$

Methods

Diversity-estimator methods:

1. Parametric
2. Non-parametric
3. Coverage based

Parametric Estimation

- Based on specific assumptions about the probability distributions of species densities
- Maximize the Likelihood of the observed f_k as a function of S and the parameters of the probability distributions of species densities.

Statistical inference from frequency count data

- S classes/taxa/species/OTUs in population.
- Assumption: Each species independently contributes Poisson-distributed number of representatives to the sample:
  - Sample: $X_1, X_2, X_3, X_4, \ldots, X_S$
- Counts ˜ zero-truncated mixed Poisson:

$$X_1 \sim Poisson(\lambda_1),\ X_2 \sim Poisson(\lambda_2),$$
$$Poisson(\lambda_3),\ \ldots,\ X_S \sim Poisson(\lambda_S)$$

The mixed-Poisson model

- Species (taxon) i contributes a Poisson-distributed number X_i of replicates to the sample – i.e., taxon i appears in the sample X_i
- Units appear independently in the sample.
- Fundamental problem: heterogeneity, i.e., unequal Poisson means λ_i
- standard approach: model λ_i‘s as i.i.d. replicates from some mixing distribution F
- Frequency counts f_k are then marginally i.i.d. F-mixed Poisson random variables Zero-truncated since zero counts X_i are unobservable

The mixed-Poisson model (Continued)

- Mixing distribution F, i.e., distribution of sampling intensities λ, is also called *species abundance distribution* (SAD) or *taxon abundance distribution* (TAD)
- Assumptions: Each species contribution to the sample is *independent and identically distributed*
- Both assumptions are probably wrong

Taxon/species abundance (mixing) distribution

- $T\left(\lambda|\boldsymbol{\theta}\right)$ – Normalised TAD, where
- $\boldsymbol{\theta}$ is vector of parameters
  - λ taxon abundance
- Quince, Curtis, and Sloan (2008)
  - Assumption: probability that individual sampled (with replacement) is from given taxon is $\frac{\lambda}{N}$, where N is size of population
  - Number of times a taxon appears in the sample will be approximately ∼Poisson($\frac{\lambda L}{N}$) , where L is the sample size and $\frac{L}{N}$ is sampling frequency

Mixing distributions F (TADs)

- Parametric, low-dimensional parameter vector
  1. **Lognormal**
  2. Inverse Gaussian
  3. Generalized inverse Gaussian (Sichel)
  4. Log-t
  5. None $\equiv$ point mass at $\lambda \equiv$ all equal species sizes Gamma (Fisher, 1943)
  6. Pareto
  7. Stable
- Finite mixture of exponentials - semiparametric

Probablity to observe a taxon

- Approach based on paper of Quince, Curtis, and Sloan (2008)
- Probability that we will observe a taxon n times:

$$P_n(r,\theta) = \int_0^\infty \frac{e^{-r\lambda}}{n!} (r\lambda)^n \, T(\lambda|\boldsymbol{\theta}) \, d\lambda, \qquad (1)$$

- $T(\lambda|\boldsymbol{\theta})$is taxon abundance (mixing) distribution
- $r = \frac{L}{N}$ is sampling ratio
- N is total population number
- θ is vector of parameters
- λ is taxon abundance

Probability with rescaled parameters

- Substitution $x = r\lambda$

$$P_n\left(\boldsymbol{\theta'}\right) = \int\limits_0^\infty \frac{e^{-x}}{n!} x^n T\left(\frac{x}{r}|\boldsymbol{\theta}\right) dx,$$

using invariance of most abundance distribution to rescaling

$$T\left(\frac{X}{r}|\theta\right) = T\left(X|\theta'\right)$$

where $\boldsymbol{\theta'}$ are rescaled parameters

$$P_n\left(\boldsymbol{\theta'}\right) = \int\limits_0^\infty \frac{e^{-x}}{n!} x^n T\left(X|\boldsymbol{\theta'}\right) dx$$

Likelihood-Multinomial distribution

$$P\left(f|\boldsymbol{\theta'},S\right) = P_0^{S-D}\prod_{i=1}^{L}\frac{P_i^{f_i}}{f_i}\frac{S!}{(S-D)!}$$

- i – abundance
- $P_0 = P(n=0|\boldsymbol{\theta'})$
- f_i – number of taxa with abundance i
- S – total number of taxa in the community
- $D = \sum_{i=1}^{L} f_i$ – number of observed taxa in sample
- $f_0 = S - D$ – number of unobserved taxa in sample

Estimating parameters of TAD (lognormal distribution)

- Estimated parameters: #taxa in community, mean and variance of lognormal distribution
- TAD as lognormal distribution numerical integration to calculate P_n
- Bayesian parameter estimation using Metropolis-Hastings MCMC with quasi-noninformative priors
- Run length 400 000 to 1 200 000 steps with 180 000 burn-in period

Estimating sampling effort with log-normal TAD 26

- From community abundance λ to sample abundance $x = r\lambda$
- For $ln\left(\lambda = X\frac{N}{L}\right) \sim N\left(\mu, \sigma^2\right)$, $ln\left(X = \lambda\frac{L}{N}\right) \sim N\left(M, V\right)$, since $M = \mu - ln\left(\frac{L}{N}\right)$
- Using the fact that $M^{new} = ln\left(\frac{L^{new}}{L}\right) + M$.
- Observed fraction of taxa $c^* = \frac{E[D]}{S} = 1 - P_0(M^{new}, V)$
- For chosen c^*, observed L and sample of parameters M and V from posterior distribution
  1. Find M^{new} s.t. $c^* = 1 - P_0(M^{new}, v)$
  2. Calculate $L^{new} = exp\left(M^{new} - M\right)$

Non-parametric estimation

- Depend on no assumptions about the probability distributions of species densities e.g.
- Chao estimator 1 (Chao 1984):

$$\hat{S}_{min} = D + \hat{E}[f_0], \text{ where}$$

$$\hat{E}[f_0] = \frac{(f_1)^2}{2f_2}$$

and f_i is number of species with abundance i.

- First order Jackknife (Burnham and Overton 1979):

$$\hat{E}\left[f_0\right] = \frac{L-1}{L} f_1,$$

where L- number of sampled units

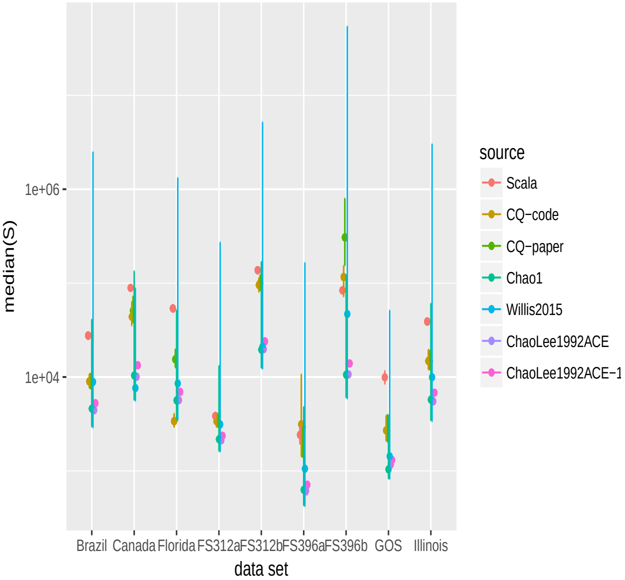
Coverage-based Estimation

- Coverage is the sum of the proportions of total density accounted for by all species encountered in the sample.
- If all species had equal density
  - $c = \frac{D}{S}$ and therefore $\hat{S} = \frac{D}{\widehat{c}}$
- Chao (Chao 1987, Chao and Lee 1992) has developed coverage-based estimators (**ACE**) by for the general case of unequal densities based on the coverage of infrequent species

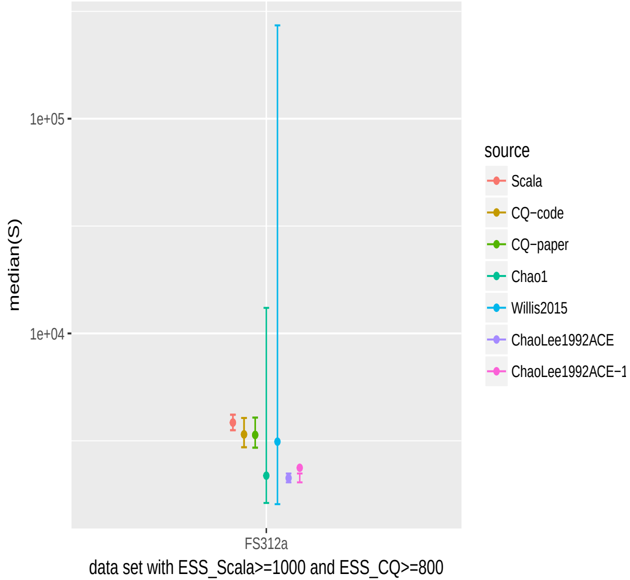
breakaway: Estimating species richness based on ratios of frequency counts

- (Willis & Bunge 2015)
- Idea:
- ratios $r\left(j\right)=\frac{(j+1)f_{j+1}}{f_j}$ are ˜ linear :
  - $r\left(j\right)=\frac{(j+1)f_{j+1}}{f_j}=\alpha+\beta_j$
- Project line downward to obtain $f_0=\#$ of unobserved species

Estimation of number of species S in community for various microbial datasets



Estimation of number of species in community for selected dataset



Summary

- Recent results look promising - more testing on real and synthetic data from various environments needed
- Future work
  - Fitting other TAD distributions: inverse Gaussian, log-Student's t, Sichel
  - Model comparison
  - Including into EBI pipeline

This is joint work with Darren Wilkinson and Tom Curtis , funded jointly with the EBI by the BBSRC BBR grant: "EBI Metagenomics Portal" led by Rob Finn at the EBI