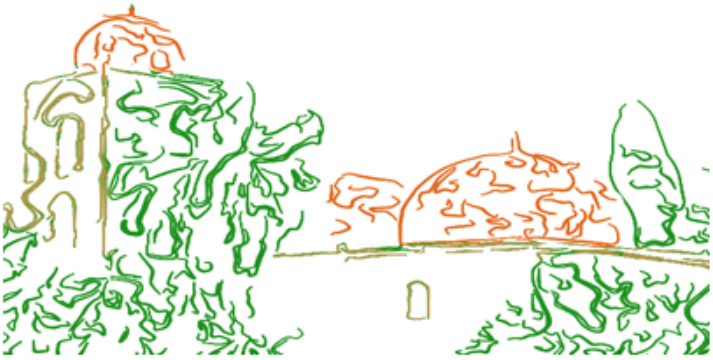
PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023



GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
-
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Sponsored by:

Città di Palermo
ift
Stazione Zoologica Anton Dohrn Napoli
Springer
Istituto Nazionale di Geofisica e Vulcanologia
Istituto Zooprofilattico Sperimentale della Sicilia "A. Mirri"
ARPA Sicilia
National Biodiversity Future Center

Contents

Modeling linkage errors in species diversity estimates: an ABC approach

D. Di Cecco[1,*] and A. Tancredi[1]

[1] *Department Memotef, University of Rome La Sapienza; davide.dicecco@uniroma1.it, andrea.tancredi@uniroma1.it*
[*]*Corresponding author*

Abstract. *The estimation of species diversity of ecological communities relies on surveying species abundances, that is, counting the number of units by species in a sample. Diversity estimators are particularly sensitive to rare species, that is, to low abundance cases. In microbial studies, rare species, in particular singletons, often represent the vast majority of the specimens in a sample. Many studies hypothesize the spurious nature of these cases, and various methodological contributions focus on estimating and eliminating the spurious singletons to avoid a gross overestimation of the total diversity of a community. We present a different approach that treats the spurious singletons as the result of false negative errors in the clustering step of the RNA sequencing. We demonstrate that the estimation of the total number of species under our scenario is equivalent to that one can obtain by discarding spurious cases. On the converse, diversity as measured by Shannon's index for example, can differ considerably. The computation of such index requires to estimate all true abundances counts, which appears to be computationally challenging. We then propose a likelihood–free Bayesian approach to the problem.*

Keywords. *Microbial Diversity; Sequencing Errors; Linkage Errors; Approximate Bayesian Computation.*

1 Introduction

The development in recent years of next generation high-throughput sequencing technology has provided the capability of analyzing an unprecedented number of DNA and RNA sequences. This technology has greatly benefited the study of microbial communities. In microbial diversity studies, environmental samples are processed in order to detect, amplify and sequence RNA genomes. The sequences are clustered into distinct species (or Operational Taxonomic Units) according to their similarities in specific genomic regions. The species are then counted by abundance, (i.e., by the number of specimens representing it). Those counts are utilized to estimate the microbial community diversity. The main interest is in estimating the "total diversity", that is, the total number of distinct species, captured or not, and in evaluating the heterogeneity of the species structure of the community. For the latter purpose, Hill's numbers are a common choice, representing a family of indexes with different degrees of sensitivity to the species abundances. Shannon's and Simpson's indexes represent particular cases. This kind of analysis can provide an indication of an ecosystem health; for example, in analysing the human gut microbiota, one can measure and compare the richness of the same community in different occasions (e.g., before and after an antibiotic treatment).

The species problem in the microbial context appears to have some peculiarities. While in macroecological studies rare species represent a small portion of the total abundance, microbial datasets are characterized by a large number of low abundance cases. In particular, singletons, that is, species represented by one specimen in the sample, often constitute the vast majority (up to 80%) of the number of observed species.

Several authors seem to agree on the fictitious nature of many rare species in microbial diversity studies (see, e.g., [8]). Even if there is no clear consensus on the frequency of such errors, or on the way their distribution is associated to the sample composition, their presence is confirmed in several ways. Experiments on samples from population with known diversity, both simulations with mock databases, and communities cultivated in vivo, confirmed an important overcount of the rare species, with consequent overestimation of the observed and total diversity.

An error in any phase of the bioinformatics pipeline can produce spurious singletons [7]. Errors in RNA sequencing produce artefactual sequences (known as chimeric sequences). Those sequences will constitute, in general, novel species which cannot be matched with any other sequence. A chimera removal step is included in many of the widely used pipelines [4]. Nonetheless, apparently, their presence is confirmed even in various curated databases of S16 sequences [5].

Since the complete removal of fictitious sequences in a pre-analysis step does not appear to be feasible at the moment, various methodological contributions aims at estimating and removing the spurious cases count (in particular singletons). See, for example, [2], [10], [11], [3]. We will call this approach where we remove units (possibly) affected by error, "discounting". This approach is consistent for a model where spurious singletons are added to the baseline counting distribution of the true abundances. This model then results in a mixture of a counting distribution (truncated in zero as the number of uncaptured species has to be estimated), and a Dirac's measure modeling the spurious singletons. Various authors propose to completely ignore the observed singletons and base the estimation solely on the other (supposedly error–free) counts.

We believe that the nature of the spurious cases is best described by linkage errors. That is, we assume that random errors occurring in sequencing result in the impossibility of a correct classification of the specimen, which cannot be associated to the correct existing species. Therefore, we can describe these cases as false negative linkage errors (or missing links), which are added to the true singletons. This approach implies a re–estimation of the "real" frequency counts for all the abundances, not just the singletons. We found that treating the excess of singletons in this way leads to significant differences in the diversity estimates with respect to the discounting approach.

In this work we focus on a secondary perspective to the linkage problem, that is, we assume the (not uncommon) condition of not having access to the actual record linkage process. Instead, we admit the possibility that our observed counts data are affected by linkage errors, and include them in our estimation process. Modeling record linkage errors in this secondary setting, appears as a computationally formidable task. We fix some simplifying assumptions on the type of error in order to tackle the issue. In particular, we assume that

- we just have missing links and no false positive record linkage errors;
- missing links create additional singletons only;
- each captured specimen has the same probability of being mistakenly classified as a singleton independently of the other.

Despite these assumptions, we resorted to a Bayesian likelihood–free approach as the most convenient method in order to estimate the true observed abundances.

2 The missing links model

Say we observed n species in our sample with abundances $y_1,...,y_n$. Let n_j be the number of species with j observed captures, such that $\sum_{j\geq 1} n_j = n$. We denote as D the observed data $(n_1,n_2,...)$. We assume that these counts can be affected by linkage errors, in particular, a portion of the singletons are due to missing links errors, and the number of true distinct captured species n^* is a portion of n. Let n_j^* be the number of true species with j captures, $j \geq 0$, and let N^* be the total number of distinct species (captured or not), $N^* = \sum_{j\geq 0} n_j^*$. Let X_i^* be the latent true number of times species i has been captured in the sample, $X_i^* \geq 0$, $i = 1,...,N^*$. The generating process of our missing links model is the following: each specimen has the same probability μ of being missclassified as a singleton. Then, for each species i captured X_i^* times we have M_i missing links, such that the registered abundance for species i is reduced from X_i^* to $X_i = X_i^* - M_i$. Then, M_i has the following Binomial conditional distribution:

$$P(M_i = m_i \,|\, X_i^* = x_i^*) = P(X_i = x_i^* - m_i \,|\, X_i^* = x_i^*) = \binom{x_i^*}{m_i} \mu^{m_i} (1-\mu)^{x_i^* - m_i}. \tag{1}$$

Let us denote as $f^*(\theta)$ the baseline counting distribution of the true captures X_i^*, defined up to a parameter θ: $\{f_j^*(\theta)\}_{j=0,1,...} = \{P(X^* = j)\}_{j=0,1,...}$. Clearly, a missing links model is completely defined by $f^*(\theta)$ and the probability μ.

In order to estimate all counts $n_0^*, n_1^*, n_2^*, ...$, we adopt a Bayesian likelihood–free approach exploiting the generating process of our model described above. We first adopt an ABC rejection sampling and then a sequential ABC to accelerate the estimation procedure as described in [6].

Mixtures of Poisson are a popular parametric choice for the baseline $f^*(\theta)$. For the sake of simplicity, in the present work we just consider Poisson and Geometric families for the baseline distribution $f^*(\theta)$. The prior setting is the following: We set $\pi(\theta)$ as the appropriate conjugate prior depending on f^*: $\lambda \sim Gamma(\alpha_\lambda, \beta_\lambda)$ for a Poisson baseline of parameter λ, and $p \sim Beta(\alpha_p, \beta_p))$ for a Geometric of parameter p. For the linkage error probability, we set $\mu \sim Beta(\alpha_\mu, \beta_\mu)$. Finally, we set an improper prior over N^* in the family $\pi(N^*) \propto 1/(N^*)^k$ (see, e.g., [1] for a sensible choice of k). All priors are independent of one another.

2.1 Rejection ABC algorithm

Note that the total number of captured specimens in our sample, call it s, remains unaltered under a missing links model. That is, we have

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n^*} x_i^* = \sum_{j\geq 1} j n_j = \sum_{j\geq 1} j n_j^* = s.$$

As a consequence, while the ABC rejection algorithm, in its simplest form, utilizes draws of the parameters from the (independent) priors, in order to exploit all the available information, we want to generate values of the parameters conditionally on the observed value s. That is, we adopt the following scheme:

1. generate values for (θ, N^*) given s and the priors $\pi(\theta)$ and $\pi(N^*)$
2. generate values $(n_0^*, n_1^*, n_2^*, ...)$ conditional on N^*, θ and s
3. generate a value for μ from the prior $\pi(\mu)$
4. generate values $\widetilde{D} = (n_0^*, \tilde{n}_1, \tilde{n}_2, ...)$ by simulating missing links over $(n_0^*, n_1^*, n_2^*, ...)$ given μ
5. retain the current generated values if a measure of distance ρ between the generated data $\widetilde{D}$ and the observed data D is below a certain threshold ε:

$$\rho(\widetilde{D}, D) < \varepsilon.$$

In the first step, under a Poisson or a Geometric baseline distribution, we can derive the analytical form of $P(N^* \,|\, s, \pi(\theta), \pi(N^*))$ and generate values of N^* accordingly. (N^*, s) constitute a sufficient statistic for the Poisson and the Geometric. Hence, the posterior distribution $\pi(\theta \,|\, N^*, s)$ is easily derived and values of θ can be easily sampled.

In step 2., we note that the distribution of $(n_0^*, n_1^*, n_2^*, ...)$ conditional on N^* and s is independent of θ. In fact, the joint distribution of N^* independent Poisson having fixed sum s is Multinomial with constant probabilities:

$$P((x_1^*, ..., x_{N^*}^*) \mid N^*, s) = Mult\big(s, (1/N^*, ..., 1/N^*)\big),$$

and, consequently,

$$P((n_0^*, n_1^*, ..., n_s^*) \mid N^*, s) = \binom{N}{n_0^* \dots n_s^*} \binom{s}{y_1^* \dots y_N^*} \frac{1}{(N^*)^s}.$$

Then we can easily generate values for $(n_0^*, n_1^*, n_2^*, ...)$. Similarly, under a Geometric assumption, all vectors $(x_1^*, ..., x_{N^*}^*)$ having fixed sum s have the same probability regardless of p, which is then equal to the reciprocal of the number of possible nonnegative integer N^*-vectors summing to s (or weak N^*–compositions of s): $\binom{N^*+s-1}{s}^{-1}$, that is,

$$P((n_0^*, n_1^*, ..., n_s^*) \,|\, N^*, s) = \binom{N^*}{n_0^* \dots n_s^*} \binom{N+s-1}{s}^{-1}.$$

As a consequence, we can generate $(x_1^*, ..., x_{N^*}^*)$ with fixed sum s with an algorithm to generate a random compositions, (see, e.g., [9]).

In step 3. we simply generate from the prior as μ is independent of N^* and s. In step 4., we simply generate missing links at random according to the binomial described in (1), modify accordingly the observed counts, and increment the number of singletons. (Note that the missing links mechanism has no effect on n_0^*). In step 5. we use the simple euclidean distance.

3 The effect of missing links on diversity

To illustrate the effect of (ignoring) a missing links mechanism on the estimation of diversity, we show the results of a simulation. As a measure of diversity we consider Shannon's diversity H (see, e.g., [3]) calculated as:

$$H = \exp\left(- \sum_{j \geq 1} n_j \frac{j}{s} \ln \frac{j}{s} \right). \tag{2}$$

We fix a scenario with $N^* = 10000$ true species and a baseline Poisson distribution f^* of parameter $\lambda = 3$. We simulated a million datasets $(n_1^*, n_2^*, \ldots)$ from the baseline distribution, then, for each such sample, we generated 30 datasets $(n_1, n_2, \ldots)$ by simulating the effect of missing links errors for 30 different values of μ ranging in $[0, 0.25]$. We calculated (2) for the original data and for the data affected by missing links. In addition, we considered the discounting procedure presented in [3] where the spurious singletons are eliminated from the observed data. The true number of singletons is estimated on the basis of this formula:

$$\widetilde{n}_1 = \frac{2n_2^2}{3n_3} + 2n_2\left(\frac{2n_2}{3n_3} - \frac{n_3}{4n_4}\right), \tag{3}$$

and Shannon's diversity is calculated over the "adjusted" data $(\widetilde{n}_1, n_2, n_3, \ldots)$.

Figure 1 summarizes the simulation results. The horizontal line indicates the average value of H (8332) over the baseline datasets $(n_1^*, n_2^*, \ldots)$. The continuous curve going upward represents the average values of H based on the simulated $n_1, n_2, \ldots$, that is, what we would obtain by ignoring any one–inflation mechanism. The continuous curve going downward represents the average values of H based on the adjusted data obtained by substituting the value n_1 with (3). That is, what we would obtain by assuming a discounting approach. The gray areas limited by red dotted lines represent the 95% confidence intervals.

Our Bayesian approach implies the generation of all counts $(n_1^*, n_2^*, \ldots)$ at each iteration of the algorithm. This allows us to estimate with ease the expectation of H under the missing links model. However, the computational complexity of the ABC approach does not allow the same number of simulations presented above. Then, we replicated the simulation 20 times with $\mu = 0.1$ and 20 times with $\mu = 0.2$ with 5000 generations for each ABC. To obtain 5000 accepted generations, each replication of the ABC required about 50 millions tries. The results were quite encouraging, producing an average of H equal to 8290 for $\mu = 0.1$ and 8420 for $\mu = 0.2$.

In conclusion, we believe that our error model represents a sensible hypothesis for the generation of additional singletons. Ignoring an existing one–inflating mechanism of this kind, implies a severe overestimation of the diversity. A discounting approach reduce sensibly the error one commits, but still leads to different results than what can be achieved with an ABC simulating the actual generating process.

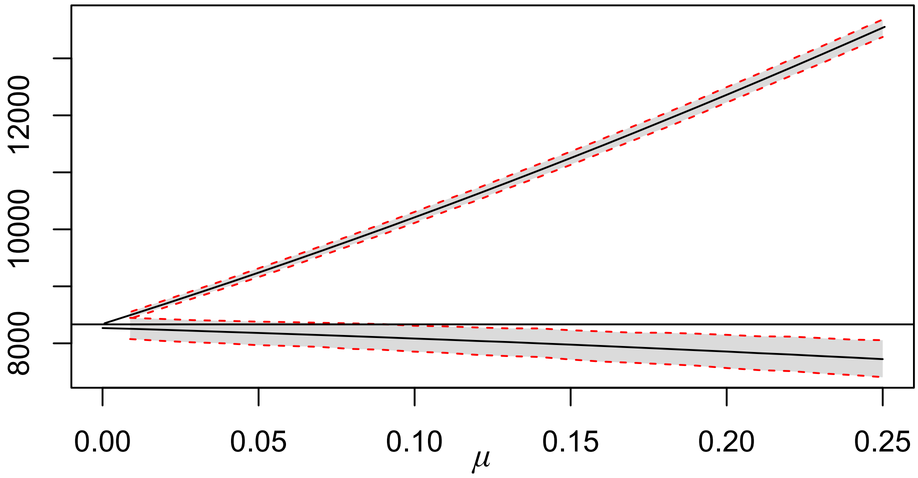


Figure 1: Shannon diversity over simulated data under different values of missing link probability μ ranging in [0,0.25].

References

[1] Barger, K., Bunge, J. (2010). Objective Bayesian estimation for the number of species. *Bayesian Analysis*, **5(4)**, 765–785.

[2] Bunge, J., Böhning, D., Allen, H., Foster, J. A. (2012). Estimating population diversity with unreliable low frequency counts. In *Biocomputing 2012: Proceedings of the Pacific Symposium* (pp. 203–212). World Sci. Publ. Hackensack, NJ.

[3] Chiu, C. H., Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, **4**, e1634.

[4] Edgar, R. (2016). Unoise2: improved error-correction for illumina 16s and its amplicon sequencing. *BioRxiv*, page 081257.

[5] Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S., Sodergren, S., et al. (2011) Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced pcr amplicons. *Genome research*, **21(3)**, 494–504.

[6] Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and computing*, **22(6)**, 1167–1180.

[7] Porter, T. M., Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular ecology*, **27(2)**, 313–338.

[8] Quince, C., Lanzen, A., Davenport, R. J., Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, **12(1)**, 1–18.

[9] Stojmenović, I. (1992). On random and adaptive parallel generation of combinatorial objects. *International journal of computer mathematics*, **42(3–4)**, 125–135.

[10] Willis, A., Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71(4)**, 1042–1049.

[11] Willis, A. (2016). Species richness estimation with high diversity but spurious singletons. *arXiv preprint arXiv:1604.02598*.