OPEN ACCESS Freely available online

PLoS one

Extrapolation of Urn Models via Poissonization: Accurate Measurements of the Microbial Unknown

Manuel E. Lladser[1]*, Raúl Gouet[2], Jens Reeder[3]

1 Department of Applied Mathematics, University of Colorado, Boulder, Colorado, United States of America, **2** Centro de Modelamiento Matemático (CNRS UMI 2807), Universidad de Chile, Santiago, Chile, **3** Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, United States of America

Abstract

The availability of high-throughput parallel methods for sequencing microbial communities is increasing our knowledge of the microbial world at an unprecedented rate. Though most attention has focused on determining lower-bounds on the α-diversity i.e. the total number of different species present in the environment, tight bounds on this quantity may be highly uncertain because a small fraction of the environment could be composed of a vast number of different species. To better assess what remains unknown, we propose instead to predict the fraction of the environment that belongs to unsampled classes. Modeling samples as draws with replacement of colored balls from an urn with an unknown composition, and under the sole assumption that there are still undiscovered species, we show that conditionally unbiased predictors and exact prediction intervals (of constant length in logarithmic scale) are possible for the fraction of the environment that belongs to unsampled classes. Our predictions are based on a Poissonization argument, which we have implemented in what we call the Embedding algorithm. In fixed i.e. non-randomized sample sizes, the algorithm leads to very accurate predictions on a sub-sample of the original sample. We quantify the effect of fixed sample sizes on our prediction intervals and test our methods and others found in the literature against simulated environments, which we devise taking into account datasets from a human-gut and -hand microbiota. Our methodology applies to any dataset that can be conceptualized as a sample with replacement from an urn. In particular, it could be applied, for example, to quantify the proportion of all the unseen solutions to a binding site problem in a random RNA pool, or to reassess the surveillance of a certain terrorist group, predicting the conditional probability that it deploys a new tactic in a next attack.

Citation: Lladser ME, Gouet R, Reeder J (2011) Extrapolation of Urn Models via Poissonization: Accurate Measurements of the Microbial Unknown. PLoS ONE 6(6): e21105. doi:10.1371/journal.pone.0021105

Editor: Dongxiao Zhu, University of New Orleans, United States of America

Received February 15, 2011; **Accepted** May 19, 2011; **Published** June 28, 2011

Funding: M. Lladser was partially supported by NASA ROSES NNX08AP60G and NIH R01 HG004872. R. Gouet was supported by FONDAP, BASAL-CMM and Fondecyt 1090216. J. Reeder was supported by a post-doctoral scholarship from the German Academic Exchange Service (DAAD). M. Lladser and R. Gouet are thankful to the project Núcleo Milenio Información y Aleatoriedad, which facilitated a research visit to collaborate in person. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: manuel.lladser@colorado.edu

Introduction

A fundamental problem in microbial ecology is the "rare biosphere" [1] i.e. the vast number of low-abundance species in any sample. However, because most species in a given sample are rare, estimating their total number i.e. α-diversity is a difficult task [2,3], and of dubious utility [4,5]. Although parametric and non-parametric methods for species estimation show some promise [6,7], microbial communities may not yet have been sufficiently deeply sampled [8] to test the suitability of the models or fit their parameters. For instance, human-skin communities demonstrate an unprecedented diversity within and across skin locations of same individuals, with marked differences between specimens [9].

In an environment composed of various but an unknown number of species, let $p_i \geq 0$ be the proportion in which a certain species i occurs. Samples from microbial communities may be conceptualized as sampling–with replacement–different colored balls from an urn. The urn represents the environment where samples are taken: soil, gut, skin, etc. The balls represent the different members of the microbial community, and each color is a uniquely defined operational taxonomic unit.

In the non-parametric setting, the urn is composed by an unknown number of colors occurring in unknown relative proportions. In this setting, the α-diversity of the urn [10] corresponds to the cardinality of the set $\{i : p_i > 0\}$. Although various lower-confidence bounds for this parameter have been proposed in the literature [11–14], tight lower-bounds on α-diversity are difficult in the non-parametric setting because a small fraction of the urn could be composed by a vast number of different colors [15]. Motivated by this, we shift our interest to predicting instead the fraction of balls with a color unrepresented in the first n observations from the urn. This is the unobservable random variable:

$$U_n = \sum_{i \notin \{X_1,\ldots,X_n\}} p_i = 1 - \sum_{i \in \{X_1,\ldots,X_n\}} p_i,$$

where $X_1, \ldots, X_n$ denote the sequence of colors observed when sampling n balls from the urn. Notice how U_n depends both on the specific colors observed in the sample, and the unknown proportions of these colors in the urn. This quantity is very useful to assess what remains unknown in the urn. For instance, the probability of discovering a new color with one additional

observation is precisely U_n, and the mean number of additional observations to discover a new color is $1/U_n$. We note that $(1-U_n)$ corresponds to what is called the conditional coverage of a sample of size n in the literature. For this reason, we refer to U_n as the *conditional uncovered probability* of the sample.

The expected value of U_n is given by:

$$u_n = \mathbb{E}(U_n) = \sum_i p_i(1-p_i)^n.$$

Unlike the conditional uncovered probability of the sample, u_n is a parameter that depends on the unknown urn composition but not on the specific colors observed in the sample. Interest in the above quantities or related ones has ranged from estimating the probability distribution of the keys used in the *Kenngruppenbuch* (the Enigma cipher book) in World War II [16], to assessing the confidence that an iterative procedure with a random start has found the global maximum of a given function [17], to predicting the probability of discovering a new gene by sequencing additional clones from a cDNA library [18]. We note that $(1-u_n)$ is called the *expected coverage of the sample* in the literature.

Various predictors of U_n and estimators of u_n have been proposed in the literature. These are mostly based on a user-defined parameter $r \geq 0$ and the statistics $N(k,n+r)$, $k=0,\ldots,r$; defined as the number of colors observed k-times, when r additional balls are sampled from the urn.

Turing and Good [19] proposed to estimate u_n using the biased statistic $v_{n,0} = N(1,n)/n$. Posteriorly, Robbins [20] proposed to predict U_n using

$$v_{n,1} = \frac{N(1,n+1)}{(n+1)}, \tag{1}$$

which he showed to be unbiased for u_n and to satisfy the inequality $\mathbb{E}\{(U_n - v_{n,1})^2\} < 1/(n+1)$. Despite the possibly small quadratic variation distance between U_n and Robbins' estimator, and as illustrated by the plots on the left side of Fig. 1, when using Robbins' estimator to predict U_n sequentially with n (to assess the quality of the predictions at various depths in the sample), we observe that unusually small or large values of $N(1,n+1)$ may offset subsequent predictions of U_n. In fact, as seen on the right-hand plots of the same figure, an offset prediction is usually followed by another offset prediction of the same order of magnitude, even 100 observations later (correlation coefficient of green clouds, $R=0.934991$ and $R=0.948600$ on top- and bottom-right plots).

Subsequently, for each $r \geq 1$, Starr [21] proposed to predict U_n using

$$v_{n,r} = \sum_{k=1}^{r} \frac{\binom{r-1}{k-1}}{\binom{n+r}{k}} \cdot N(k,n+r). \tag{2}$$

Even though $v_{n,r}$ is the minimum variance unbiased estimator of u_n based on r additional observations from the urn [22], Starr showed that $v_{n,r}$ may be strongly negatively correlated with U_n when $r=1$ (note that Starr's and Robbins' estimators are identical when $r=1$). Furthermore, the sequential prediction of U_n via Starr's estimator is affected by issues similar to Robbins' estimator, which is also illustrated in Fig. 1, even when the parameter r is set as large as possible, namely $(n+r)$ is equal to the sample size (correlation coefficient of orange clouds, $R=0.996407$ and $R=0.984397$ on top- and bottom-right, respectively). We observe that $v_{n,1}$ and $v_{n,r}$ are indistinguishable in a linear scale when $r \ll n$ because, for each $n,r \geq 1$, it applies that (see Materials and Methods):

$$|v_{n,1} - v_{n,r}| \leq \frac{2(r-1)}{n+1} + \frac{r-1}{r+n-1}. \tag{3}$$

In terms of prediction intervals, if z_α denotes the α upper quantile of a standard Normal distribution, it follows from Esty's analysis [23] that if $N(1,n)/n$ is not very near 0 or 1 then

$$v_{n,0} \pm z_{\alpha/2} \cdot \sqrt{\frac{v_{n,0}(1-v_{n,0})}{n} + 2\frac{N(2,n)}{n^2}}, \tag{4}$$

is approximately a $100(1-\alpha)\%$ prediction interval for U_n. In practice, and as seen in Fig. 2, when the center of the interval is of a similar or lesser order of magnitude than its radius, the ratio between the upper- and lower-bound of these intervals may oscillate erratically, sometimes over several orders of magnitude. This can be an issue in assessing the depth of sampling in rich environments. For instance, to be highly confident that $10^{-5} \leq U_n \leq 10^{-3}$ is not of practical use because one may need from 1000 to 100,000 additional observations to discover a new species.

The issues of the aforementioned methods are somewhat expected. On one hand, the problem of predicting U_n is very different from estimating u_n: the former requires predicting the exact proportion of balls in the urn with colors outside the random set $\{X_1,\ldots,X_n\}$, rather than in average over all possible such sets. On the other hand, the point estimators of u_n are unlikely to predict U_n accurately in a logarithmic scale, unless the standard deviation of U_n is small relative to U_n. Finally, the methods we have described from the literature were designed for static situations i.e. to predict U_n or estimate u_n when n is fixed.

Results

Embedding Algorithm

Here we propose a new methodology to address the issues of the methods presented in the Introduction to predict U_n. Our methodology lends itself better for a sequential analysis and accurate predictions in a logarithmic scale; in particular, also in a linear scale–though it relies on randomized sample sizes. Due to this, in static situations i.e. for fixed sample sizes, our method only yields predictions for a random sub-sample of the original sample.

Randomized sample sizes are more than just an artifact of our procedure: due to Theorem 1 below, for any predetermined sample size, there is no deterministic algorithm to predict U_n and $\ln(U_n)$ unbiasedly, unless the urn is composed by a known and flat distribution of colors. See the Materials and Methods section for the proofs of our theorems.

Theorem 1 *If $f : [0,1] \to [-\infty,+\infty]$ is a continuous and one-to-one function then the following two statements are equivalent: (i) there is a non-randomized algorithm based on $(X_1,\ldots,X_n)$ to predict $f(U_n)$ conditionally unbiased; (ii) the urn is composed by a known and equidistributed number of colors.*

Our methodology is based on a so called Poissonization argument [24]. This technique is often used in allocation problems to remove correlations [25]. It was applied in [26] to show that the cardinality of the random set $\{X_1,\ldots,X_n\}$ is asymptotically Gaussian after the appropriate renormalization. Mao and Lindsay [27] used implicitly a Poissonization argument to argue that

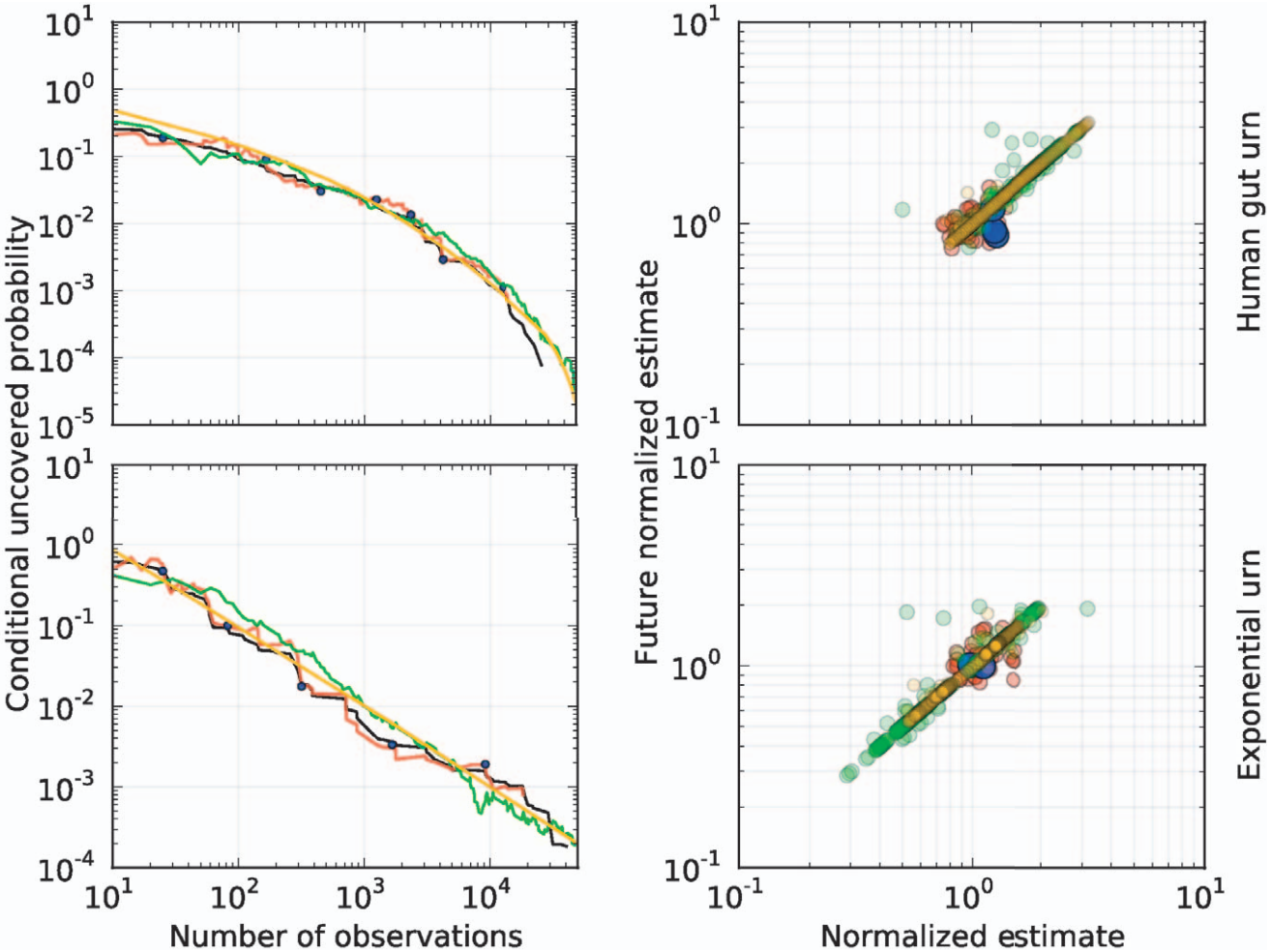

Figure 1. Point predictions in a human-gut and exponential urn. Plots associated with a human-gut (top-row) and exponential urn (bottom-row). Left-column, sequential predictions of the conditional uncovered probability (black), as a function of the number n of observations, using Robbins' estimator in equation (1) (green), Starr's estimator in equation (2) (orange), and the Embedding algorithm (blue, red), over a same sample of size 50,000 from each urn. Starr's estimator was implemented keeping $n+r=50{,}000$. Blue predictions correspond to consecutive outputs of the Embedding algorithm in Table 1, which was reiterated until exhausting the sample using the parameter $r=25$. Red predictions correspond to outputs of the algorithm each time a new species was discovered. Right-column, correlation plots associated with consecutive predictions of the conditional uncovered probability (normalized by its true value at the point of prediction), under the various methods. The green and orange clouds correspond to pairs of predictions, 100-observations apart, using Robbins' and Starr's estimators, respectively. Blue and red clouds correspond to pairs of consecutive outputs of the Embedding algorithm, following the same coloring scheme than on the left plots. Notice how the red and blue clouds are centered around (1,1), indicating the accuracy of our methodology in a log-scale. Furthermore, the green and orange clouds show a higher level of correlation than the blue and red clouds, indicating that our method recovers more easily from previously offset predictions. In each urn, our predictions used the 50,000 observations and a HPP with intensity one–simulated independently from the urn–to predict sequentially the uncovered probability of the first part of the sample. See Fig. 4 for the associated rank curve in each urn.
doi:10.1371/journal.pone.0021105.g001

intervals such as in equation (4) have a $100(1-\alpha)\%$ asymptotic confidence, under the hypothesis that the times at which each color in the urn is observed obey a homogeneous Poisson point process (HPP) with a random intensity. Here, asymptotic means that the α-diversity tends to infinity, which entails adding colors into the urn. Our approach, however, is not based on any assumption on the times the data was collected, nor on an asymptotic rescaling of the problem, but rather on the embedding of a sample from an urn into a HPP with intensity 1 in the semi-infinite interval $[0,+\infty)$. We emphasize that the HPP is a mathematical artifice simulated independently from the urn.

In what follows, $r\geq 1$ is a user-defined integer parameter. We have implemented the Poissonization argument in what we call the *Embedding algorithm* in Table 1. For a schematic description of the algorithm see Fig. 3 and, for its heuristic, consult the Materials and Methods section.

Suppose that a set I of colors is already known to belong to the urn and let $p_I=\sum_{i\in I}p_i$ be the coverage probability of the colors in this set. We note that, in the context of the previous discussion, $U_n=(1-p_I)$ with $I=\{X_1,\ldots,X_n\}$.

To predict $(1-p_I)$, draw balls from the urn until r colors outside I are observed. Visualize each observation as a colored point in the interval $[0,+\infty)$. The Poissonization consists in spacing these points out using independent exponential random variables with mean one. Due to the thinning property of Poisson point processes [28], the position T_r of the point farthest apart from 0 has a Gamma distribution with mean $r/(1-p_I)$. We may exploit this to obtain conditionally unbiased predictors and exact

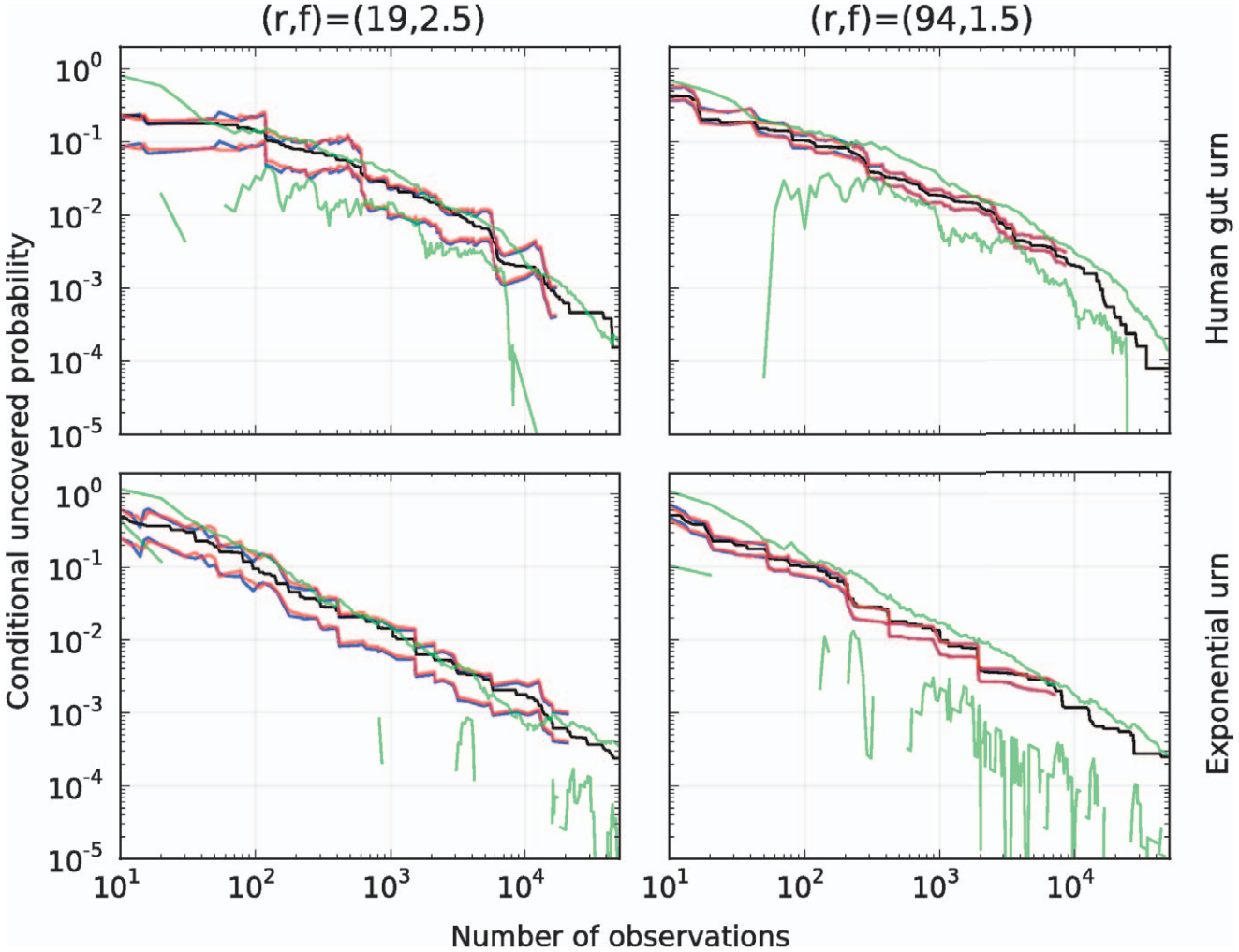

Figure 2. Prediction intervals in the human-gut and exponential urn. 95% prediction intervals for the conditional uncovered probability (black) of the human-gut and exponential urn as a function of the number of observations. Esty's prediction intervals in equation (4) (green), and predictions intervals based on the Embedding algorithm (blue, red), using the parameters $(r,f)=(19,2.5)$ and $(r,f)=(94,1.5)$ on the left and right, respectively. Blue and red curves correspond to the conservative-lower and -upper prediction intervals for the uncovered probability, respectively. The missing segments on the lower green-curves correspond to Esty's prediction intervals that contained 0. Although the upper- and lower-bound of the Esty's intervals may be of different order of magnitude, our method produces intervals of a constant length in logarithmic scale. This length is controlled by the user-defined parameter f. In each urn, our method predicted accurately the uncovered probability of a random sub-sample of the 50,000 observations from the urn. See Fig. 4 for the associated rank curve in each urn.
doi:10.1371/journal.pone.0021105.g002

prediction intervals for $(1-p_I)$ and $\ln(1-p_I)$ as follows. Regarding direct predictions of $\ln(1-p_I)$, note that measuring $(1-p_I)$ in a logarithmic rather than linear scale makes more sense when deep sampling is possible.

Theorem 2 *Conditioned on I and the event $p_I<1$, the following applies:*

(i) *If $r\geq 3$ then $(r-1)/T_r$ is unbiased for $(1-p_I)$, with variance $(1-p_I)^2/(r-2)$.*

(ii) *If $r\geq 1$ and $\gamma=0.57721566\ldots$ denotes Euler's constant then $-\ln(T_r)-\gamma+\sum_{i=1}^{r-1}1/i$ is unbiased for $\ln(1-p_I)$, with variance $\sum_{i=r}^{\infty}1/i^2$, which is bounded between $1/r$ and $1/(r-1)$.*

(iii) *If $r\geq 1$, $0<\alpha<1$ and $0\leq a\leq b\leq+\infty$ are such that*

$$\int_a^b \frac{x^{r-1}}{(r-1)!}e^{-x}dx=(1-\alpha), \qquad (5)$$

then the interval $[a/T_r,b/T_r]$ contains $(1-p_I)$ with exact probability $(1-\alpha)$; in particular, $[\ln(a/T_r),\ln(b/T_r)]$ contains $\ln(1-p_I)$ also with probability $(1-\alpha)$.

We note that $(r-1)/T_r$ is the uniformly minimum variance unbiased estimator of $(1-p_I)$ based on r exponential random variables with unknown mean $1/(1-p_I)$. Furthermore, $(1-p_I)\cdot T_r/r$ converges almost surely to 1, as r tends to infinity; in particular, the point predictors in part (i) and (ii) are strongly consistent.

We also note that the logarithm of the statistic in part (i) underestimates $\ln(1-p_I)$ in average. In fact, the difference between the natural logarithm of the statistic in (i) and the statistic in (ii) is $\gamma+\ln(r-1)-\sum_{i=1}^{r-1}1/i$, which is negative for $r\geq 2$, and increases to zero as r tends to infinity. From a computational stand point, however, the statistics $\ln((r-1)/T_r)$ and $-\ln(T_r)-\gamma+\sum_{i=1}^{r-1}1/i$ differ by at most 1%-units when $r\geq 51$. The same precision may

Table 1. Embedding algorithm.

Input:	$r \geq 1$, a set I of colors known to be in the urn, and constants $0 \leq a < b \leq +\infty$ that satisfy condition (5).
Output:	Unbiased predictor of $(1-p_I)$, $100(1-\alpha)\%$ prediction interval for $(1-p_I)$ and an updated set I of colors known to belong to the urn.
Step 1.	Assign $i := 0$, $j := 0$, and $J := I$.
Step 2.	While $j < r$ assign $i := (i+1)$, and sample with replacement a ball from the urn. Let c be the color of the sampled ball. If $c \notin I$ then assign $j := (j+1)$ and $J := J \cup \{c\}$.
Step 3.	Simulate $T_r \sim Gamma(i,1)$, and assign $I := J$.
Step 4.	Output $(r-1)/T_r$, $[a/T_r, b/T_r]$ and I.

doi:10.1371/journal.pone.0021105.t001

be reached for smaller values of r if larger bases are utilized. For instance, in base-10, the discrepancy will be at most 1% for $r \geq 23$.

In regards to part (iii) of the theorem, we note that our prediction intervals for $(1-p_I)$ cannot contain zero unless $a=0$. On the other hand, since the density function used in equation (5) is unimodal, the shortest prediction interval for $(1-p_I)$ corresponds to a pair of non-negative constants $a < (r-1) < b$ such that:

$$a^{r-1}e^{-a} = b^{r-1}e^{-b} \text{ and } \int_a^b \frac{x^{r-1}}{(r-1)!}e^{-x}dx = (1-\alpha). \quad (6)$$

Similarly, optimal prediction intervals for $\ln(1-p_I)$ follow when

$$a^{r}e^{-a} = b^{r}e^{-b} \text{ and } \int_a^b \frac{x^{r}}{r!}e^{-x}dx = (1-\alpha), \quad (7)$$

with $0 < a < r < b$ (see Materials and Methods for a numerical procedure to approximate these constants). In either case, because $\{(1-p_I)\cdot T_r - r\}/\sqrt{r}$ converges in distribution to a standard Normal as r tends to infinity, one may select in (5) the approximate constants $a = r-1-\sqrt{r-1}\cdot z_{\alpha/2}$ and $b = r-1+\sqrt{r-1}\cdot z_{\alpha/2}$. With these approximate values, if $0 < z_{\alpha/2} < \sqrt{r-1}$ then the true confidence c of the associated prediction intervals satisfies (see Materials and Methods):

$$\exp\left\{z_{\alpha/2}\cdot\sqrt{r-1} + \frac{z_{\alpha/2}^2}{2} + (r-1)\cdot\ln\left(1-\frac{z_{\alpha/2}}{\sqrt{r-1}}\right) - \frac{1}{12(r-1)}\right\} \leq \frac{c}{1-\alpha} \leq \exp\left\{\frac{z_{\alpha/2}^3}{3\sqrt{r-1}}\right\}. \quad (8)$$

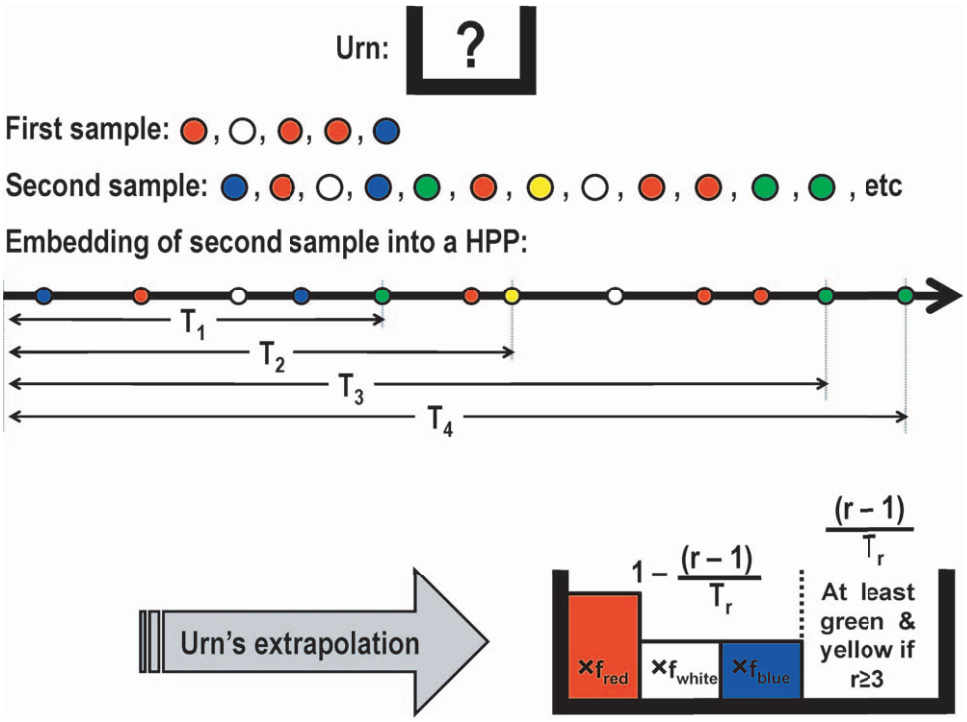

Figure 3. Schematic description of the Embedding algorithm. Suppose that in a first sample from an urn you only observe the colors red, white and blue; in particular, $I = \{\text{red,white,blue}\}$. Let m be the unknown proportion in the urn of balls colored with any of these colors i.e. $m = p_I$. To estimate $(1-m)$, sample additional balls from the urn until observing r balls with colors outside I. Embed the colors of this second sample into a homogeneous Poisson point process with intensity one; in particular, the average separation of consecutive points with colors outside I are independent exponential random variables with mean $1/(1-m)$. The unknown quantity $(1-m)$ can be now estimated from the random variable T_r. As a byproduct of our methodology, conditional on I, if f_i denotes the relative proportion of color i in the first sample then $(1-(r-1)/T_r)\times f_i$ predicts the true proportion of color i in the urn.
doi:10.1371/journal.pone.0021105.g003

(The term on the exponential on the left-hand side above is big-O of $z_{\alpha/2}^3/\sqrt{r-1}$; in particular, the lower-bound is of the same asymptotic order than the upper-bound.) We note that the constants produced by the Normal approximation may be crude for relatively large values of r, as seen in Table 2.

As high-throughput technologies allow deeper sampling of microbial communities, it will be increasingly important to have upper- and lower-bounds for $(1-p_I)$ of a comparable order of magnitude. Since the prediction intervals for this quantity in Theorem 2 are of the form $[a/T_r, b/T_r]$, and the ratio between the upper- and lower-bound of this interval is b/a, one may wish to determine constants a and b such that, not only (5) is satisfied, but also

$$b/a = f, \tag{9}$$

where $f>1$ is a user-defined parameter. Not all values of f are attainable for a given r and confidence level. In fact, the smallest attainable value is given by the constants associated with the optimal prediction interval for $\ln(1-p_I)$. Equivalently, f is attainable if and only if

$$f \geq b^*/a^*, \text{ where } \int_{a^*}^{b^*} \frac{x^r}{r!} e^{-x} dx = (1-\alpha)$$
$$\text{and } (a^*)^r e^{-a^*} = (b^*)^r e^{-b^*} \text{ with } a^* < r < b^*.$$

Conversely, and as stated in the following result, any value of $f>1$ is attainable at a given confidence level, provided that the parameter r is selected sufficiently large.

Theorem 3 *Let $0<\alpha<1$ and $f>1$ be fixed constants. For each r sufficiently large, there are constants $0<a<b<+\infty$ such that (5) and (9) are satisfied.*

For a given parameter f, there are at most two constants $0<c_1<c_2<+\infty$ such that $[c_1/T_r, f\cdot c_1/T_r]$ and $[c_2/T_r, f\cdot c_2/T_r]$ are prediction intervals for $(1-p_I)$ with exact confidence $(1-\alpha)$. We refer to these as *conservative-lower* and *conservative-upper prediction intervals*, respectively. We refer to intervals of the form $[0, c_0/T_r]$ and $[c_3/T_r, 1]$ as *upper-* and *lower-bound prediction intervals*, respectively. See Table 3 for the determination of these constants for various values of r when $\alpha=5\%$.

Effect of non-randomized sample sizes

The Embedding algorithm provides conditionally unbiased predictors and intervals for $(1-p_I)$ and $\ln(1-p_I)$, provided that an arbitrary number of additional observations is possible until observing r balls with colors outside I. When dealing with fixed sample sizes, there is a positive probability of not meeting this condition, in which case the Embedding Algorithm is inconclusive. In large samples however, such as those collected in microbial datasets, the algorithm may be applied sequentially until it yields an inconclusive prediction. In such case, the true confidence of the prediction intervals produced by the algorithm satisfy the following.

Theorem 4 *Suppose that condition (5) is satisfied. Conditioned on I, if r balls with colors outside I are observed in the next k draws from the urn, then the true confidence c of the prediction interval for $(1-p_I)$ produced by the Embedding algorithm satisfies:*

(i) if $a=0$ then $(1-\alpha) \leq c \leq (1-\alpha) + \mathbb{P}[\Gamma > b]$;

(ii) if $a>0$ then $(1-\alpha) - \mathbb{P}[N>k] \leq c \leq (1-\alpha) + \mathbb{P}[\Gamma > b]$, where Γ is a Gamma random variable with parameters $(r,1)$, and N is a Negative Binomial random variable with parameters $(r, 1-p_I)$.

Table 2. Optimal versus asymptotic 95% prediction intervals.

r	Prediction interval for	Optimal constants	Gaussian approximation	Relative error ~
30	$(1-p_I)$	$a=19.66173485$ $b=40.91013748$	$a=18.44527092$ $b=39.55472908$	-6.2% -3.3%
30	$\ln(1-p_I)$	$a=20.48229580$ $b=42.08921485$	Same as above	-9.9% -6.0%
120	$(1-p_I)$	$a=98.86695443$ $b=141.6966834$	$a=97.61931714$ $b=140.3806829$	-1.3% -0.9%
120	$\ln(1-p_I)$	$a=99.77743953$ $b=142.7861762$	Same as above	-2.2% -1.7%

doi:10.1371/journal.pone.0021105.t002

Table 3. Constants associated with 95% prediction intervals.

r	c_0	c_1	c_2	c_3
1	2.995732274	*	*	0.051293294
2	4.743864518	*	*	0.355361510
3	6.295793622	*	*	0.817691447
4	7.753656528	0.806026244	1.360288674	1.366318397
5	9.153519027	0.924031159	1.969902541	1.970149568
6	10.51303491	1.053998892	2.61300725	2.613014744
7	11.84239565	1.185086999	3.28531552	3.285315692
8	13.14811380	1.315076338	3.98082278	3.980822786
9	14.43464972	1.443547021	4.69522754	4.695227540
10	15.70521642	1.570546801	5.42540570	5.425405697
11	16.96221924	1.696229569	6.16900729	6.169007289
12	18.20751425	1.820753729	6.92421252	6.924212514
13	19.44256933	1.944257623	7.68957829	7.689578292
14	20.66856908	2.066857113	8.46393752	8.463937522
15	21.88648591	2.188648652	9.24633050	9.246330491
16	23.09712976	2.309712994	10.03595673	10.03595673
17	24.30118368	2.430118373	10.83214036	10.83214036
18	25.49923008	2.549923010	11.63430451	11.63430451
19	26.69177031	2.669177032	12.44195219	12.44195219
20	27.87923964	2.787923964	13.25465160	13.25465160
21	29.06201884	2.906201884	14.07202475	14.07202475
22	30.24044329	3.024044329	14.89373854	14.89373854
23	31.41481021	3.141481021	15.71949763	15.71949763
24	32.58538445	3.258538445	16.54903872	16.54903871
25	33.75240327	3.375240328	17.38212584	17.38212584

Constants associated with 95% upper-bound, conservative-lower, conservative-upper and lower-bound prediction intervals for $(1-p_I)$, when $1 \leq r \leq 25$ and $f=10$. By definition, this means that $\int_0^{c_0} \frac{x^{r-1}}{(r-1)!} e^{-x} dx = 0.95$ and $\int_{c_3}^{\infty} \frac{x^{r-1}}{(r-1)!} e^{-x} dx = 0.95$. Furthermore, the constants $c_1 \leq c_2$ are solutions to the equation:

$$\int_c^{10c} \frac{x^{r-1}}{(r-1)!} e^{-x} dx = 0.95, \quad c \geq 0,$$

solved numerically with Newton's method using Maple 13.02. This equation may have at most two different solutions, and star (*) denotes that the equation has no solution.

doi:10.1371/journal.pone.0021105.t003

Thus, if the Embedding algorithm produces an output in what remains of a finite sample size, the upper-bound prediction interval for $(1-p_I)$ has at least the user-defined confidence. This is perhaps the case of most interest in applications: it allows the user to estimate the least number of additional samples to observe a color not seen in any sample. For the other three interval types, the true confidence is approximately at least the targeted one if the probability that the algorithm produces an output in what remains of the sample is large.

Discussion

Comparisons with Robbins-Starr estimators

Note that, like Robbins' and Starr's estimators, our method requires extracting additional balls from the urn to make a prediction. However, unlike the methods of the Introduction, our method uses only the additionally collected data–instead of all the data ever collected from the urn–to make a prediction. In terms of sequential analysis, this is advantageous to recover from earlier erroneous predictions (we expand on this point in the next section, see Fig. 1).

In what remains of this section, $I=\{X_1,\ldots,X_n\}$ hence $(1-p_I)=U_n$, the conditional uncovered probability of a sample of size n. Furthermore, to rule out trivial cases, we assume that $U_n<1$ with positive probability i.e. the urn is composed by more than just balls of a single color.

Part (i) of Theorem 1 provides a conditionally unbiased predictor for U_n. We can show, however, that Robbins' and Starr's estimators are not conditionally unbiased for U_n in the non-parametric case when $r<n/6+1$. To see this argument, first notice that $|v_{n,r}-v_{n,1}|\leq 3(r-1)/n$ due to the inequality (3). On the other hand, if i is a color in the urn such that $p_i>0$ then

$$\mathbb{E}(v_{n,1}|I=\{i\})=\frac{1-p_i}{n+1}.$$

As a result:

$$\begin{aligned}&(1-p_i)-\mathbb{E}(v_{n,r}|I=\{i\})\\&=(1-p_i)-\mathbb{E}(v_{n,1}|I=\{i\})+\mathbb{E}(v_{n,1}-v_{n,r}|I=\{i\})\\&\geq\frac{1}{2}\left\{1-\frac{6(r-1)}{n}-p_i\right\}.\end{aligned}$$

Hence, if there exists a color i in the urn that makes the above quantity strictly positive (there are infinitely many such urns, including all urns composed by infinitely many colors, because $r<n/6+1$) then $v_{n,r}$ cannot be conditionally unbiased for U_n.

On the other hand, due to parts (i) and (ii) in Theorem 1, we obtain (see Materials and Methods):

$$\rho\left(U_n,\frac{r-1}{T_r}\right)=\sqrt{\frac{\mathbb{V}(U_n)}{\mathbb{V}((r-1)/T_r)}},\qquad(10)$$

$$\rho\left(\ln(U_n),-\ln(T_r)-\gamma+\sum_{i=1}^{r-1}\frac{1}{i}\right)=\sqrt{\frac{\mathbb{V}(\ln(U_n))}{\mathbb{V}(\ln(T_r))}},\qquad(11)$$

where ρ denotes correlation and $\mathbb{V}$ variance. Consequently, the point predictors in Theorem 1 are positively correlated with the quantities they were designed to predict. This contrasts with Robbins' estimator, which may be strongly negatively correlated with U_n. For instance, if $p_i=1/k$ for k different colors in the urn, it is shown in [21] that the asymptotic correlation between U_n and Robbins' estimator $v_{n,1}$ is asymptotically negative when n/k converges to a strictly positive but finite constant λ. In this same regime but provided that $r\ll\sqrt{n}$, we can show that (see Materials and Methods):

$$\limsup_{n\to\infty}\rho(U_n,v_{n,r})\leq\frac{\lambda\cdot e^{-\lambda}-\lambda\cdot(1+3\lambda)\cdot e^{-2\lambda}}{2\sqrt{\lambda\cdot e^{-2\lambda}-\lambda\cdot(2+\lambda^2)\cdot e^{-3\lambda}+\lambda\cdot(1+\lambda^3)\cdot e^{-4\lambda}}}.\qquad(12)$$

Since the right-hand side above is negative for all λ sufficiently small, Starr's estimator $v_{n,r}$ may also have a strong negative correlation with U_n when r is much smaller than $\sqrt{n}$.

A further calculation based on parts (i) and (ii) in Theorem 1 shows that

$$\mathbb{V}\left(\frac{r-1}{T_r}\right)=\mathbb{V}(U_n)+\frac{\mathbb{E}(U_n^2)}{r-2},$$

$$\mathbb{V}(\ln(T_r))=\mathbb{V}(\ln(U_n))+\sum_{i=r}^{\infty}\frac{1}{i^2}.$$

In particular, for fixed n, the correlations in equations (10) and (11) approach to one as r tends to infinity.

Finally, for non-trivial urns with finite α-diversity, i.e. urns composed by balls with at least two but a finite number of different colors, one can show for fixed r that the correlation in equation (10) approaches $\sqrt{(r-2)/(r-1)}$ as n tends to infinity. Furthermore, if we again assume that $p_i=1/k$ for k different colors in the urn and n/k converges to a strictly positive but finite constant, then the correlation in equation (10) approaches zero from above. As we pointed out before, in this regime, Robbins' estimator is asymptotically negatively correlated with U_n.

Selection of parameters

There are two main criteria to select the parameter r of the Embedding algorithm in a concrete application.

One criteria applies for point predictors. In this case, conditioned on I, the standard deviation of the relative error of our prediction of $(1-p_I)$ is $1/\sqrt{r-2}$ (Theorem 2, part (i)). To predict $(1-p_I)$, r should be therefore selected as small as possible so as to meet the user's tolerance on the average relative error of our predictions. The same criteria applies for point predictors of $\ln(1-p_I)$, for which the standard deviation of the absolute error is of order $1/\sqrt{r}$, uniformly for all $p_I<1$ (Theorem 2, part (ii)).

A different criteria applies for prediction intervals. In this case, conditioned on I, the user should first specify the confidence level, and how much larger he wants the upper-prediction-bound to be in relation to the lower-prediction bound of $(1-p_I)$. Since the ratio between these last two quantities is given by the parameter f in (9), r should be selected as small as possible to meet the user's pre-specified factor f for the given confidence level of the prediction interval (Theorem 3). See Table 4 for the optimal choice of r for various values of f when $\alpha=5\%$. Note that for the selected parameter r, the constants associated with the optimal prediction intervals are given in equations (6) and (7), see Materials and Methods.

Simulations on analytic and non-analytic urns

We tested our methods against an urn with an exponential relative abundance rank curve over 500 species, and an urn

Table 4. Optimal selection of parameter r in terms of parameter f.

f	r	c_1	c_2
80	2	0.0598276655	0.355361510
48	2	0.1013728884	0.355358676
40	2	0.1231379857	0.355320458
24	2	0.226833483	0.346045204
20	3	0.320984257	0.817610455
12	3	0.590243030	0.787721610
10	4	0.806026244	1.360288674
6	6	1.822307383	2.58658608
5	7	2.48303930	3.22806682
3	14	7.17185045	8.27008349
2.5	19	11.26109001	11.96814857
1.5	94	75.9077267	76.5492088
1.25	309	275.661191	275.949782

Constants associated with the controlled upper- to lower-bound ratio prediction intervals for $(1-p_I)$, when $\alpha=5\%$; in particular, for each f and r, $[c_1/T_r, f\cdot c_1/T_r]$ and $[c_2/T_r, f\cdot c_2/T_r]$ contain $(1-p_I)$ with a 95% probability. For each f, the smallest value of r for which the equation:

$$\int_c^{cf} \frac{x^{r-1}}{(r-1)!}e^{-x}dx=0.95, \quad c\geq 0;$$

admits a solution, is reported. Numerical values where determined using Maple 13.02.

doi:10.1371/journal.pone.0021105.t004

matching the observed distribution of microbes in a human-gut sample from [29]. We also analyzed a sample from a human-hand microbiota found in [30]. The gut and hand data are part of the largest microbial datasets collected thus far (see Fig. 4 for the relative abundance rank curve associated with each urn). The relative abundance rank curve, or for simplicity "rank curve", associated with an urn is a graphical representation of its composition: the height of the graph above a non-negative integer i is the fraction of balls in the urn with the i-th most dominant color.

The blue dots and red curves on the plots on the left side in Fig. 1 show very accurate point predictions in log-scale of the conditional uncovered probability (as a function of the number of observations), when we apply the Embedding algorithm to a sample of size 50,000 from the human-gut and exponential urn, respectively. In both instances, the parameter r of the Embedding algorithm was set to 25. The accuracy of our method is confirmed by the red clouds on the plots on the right side of Fig. 1, which are centered around (1,1). The red clouds also indicate that our predictions recover more easily from offset predictions as compared to Robbins' and Starr's (correlation coefficient of red clouds, $R=0.715451$ and $R=0.244014$ on top- and bottom-right, respectively). This is to be expected because the Embedding algorithm relies only on the additionally collected data to make a new prediction, whereas Robbins' and Starr's estimators use all the data ever collected from the urn. On the other hand, the red and blue curves in Fig. 2 show that the conservative-upper and -lower prediction intervals of the conditional uncovered probability (also as a function of the number of observations) contain this quantity with high probability and, unlike Esty's intervals, have a constant length in logarithmic scale. The intervals on the plots on the right side are tighter than those on the left because of the decrease of the parameter f from 2.5 to 1.5. In each case, the

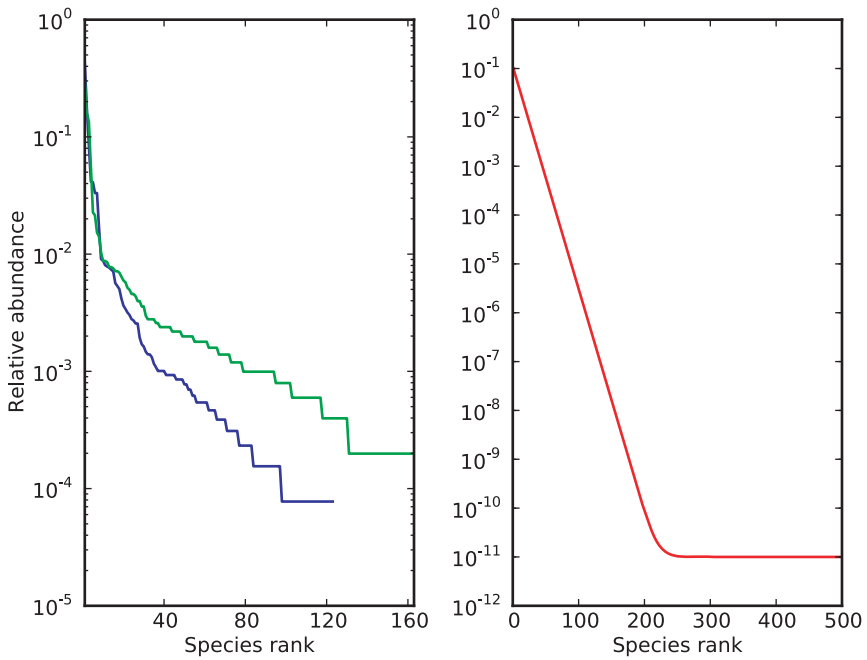

Figure 4. Rank curves associated with the human-gut, human-hand and exponential urn. In a rank curve, the relative abundance of a species is plotted against its sorted rank amongst all species, allowing for a quick overview of the evenness of a community. On the left, rank curves associated with the human-gut (blue) and -hand data (green) show a relatively small number of species with an abundance greater than 1%, and a long tail of relatively rare species. The right rank curve of the exponential urn (red) simulates an extreme environment, where relatively excessive sampling is unlikely to exhaust the pool of rare species.

doi:10.1371/journal.pone.0021105.g004

parameter r was selected according to the guidelines in Table 4. We note that sequential predictions based on the Embedding Algorithm in figures 1 and 2 were produced until the algorithm yielded inconclusive predictions. For this reason, our predictions ended before exhausting each sample.

In the human-hand dataset, 163 species were observed in a sample of size 5034. To simulate draws with replacement from this environment, we produced a random permutation of the data (see Materials and Methods section). Using the Embedding algorithm with parameters $(r,f)=(50,2)$, and according to our point predictor, 133 of the species observed in the sample represent $\sim$98.3% of that hand environment; in particular, the remaining $\sim$1.7% is composed by at least 30 species. Furthermore, according to our upper-bound prediction interval, and with at least a 95% confidence, the species not represented in the sample account for less than 2.2% of that environment.

To test the above predictions, we simulated the rare biosphere as follows. We hypothesized that our point prediction of the conditional uncovered probability could be offset by up to one order of magnitude. We also hypothesized that the number of unseen species in the sample had an exponential relative abundance rank curve, composed either by 10, 100 or 1000 species. This leads to nine different urns in which to test our methods. These urns are devised such that they gradually change from the almost unchanged urn in the bottom left corner to the urn in the upper right, which is dominated by rare species (see Fig. 5 for the associated rank curves). As seen on the plots in Fig. 6, the Embedding algorithm yields very accurate predictions in each of these nine scenarios, for all the sample sizes considered.

As seen in Fig. 7, our predictions are also in excellent agreement with the human-gut dataset when we simulate the rare biosphere. As expected, the conditional uncovered probability almost always lies between the predicted bounds. We also note that the predictions based on the Embedding algorithm are accurate even for a small number of observations. This suggests that our algorithm can be applied to deeply as well as shallowly sampled environments.

Materials and Methods

Heuristic behind the Embedding algorithm

The number of times a rare color occurs in a sample from an urn is approximately Poisson distributed. In the non-parametric setting, a direct use of this approximation is tricky because "rare" is relative to the sample size and the unknown urn composition. The embedding into a HPP is a way to accommodate for the Poisson approximation heuristic, without making additional assumptions on the urn's composition. To fix ideas, imagine that no ball in the urn is colored black. Make up a second urn with a single ball colored black. We refer to this as the "black-urn". Now sample (with replacement) balls according to the following scheme: draw a ball from the original- versus black-urn with probability ε and $(1-\varepsilon)$, respectively, where $\varepsilon>0$ is a fixed but small parameter. Under this sampling scheme, even the most abundant colors in the original-urn are rare. In particular, the smaller ε is, the closer is the distribution of the number of times a particular set of colors (excluding black) is observed to a Poisson distribution. This approach is not very practical, however, because the number of samples to observe a given number of balls from the original urn can be astronomically large when ε is very small. To overpass this issue imagine drawing a ball every ε-seconds. Draws from the original urn will then be apart $\varepsilon T_\varepsilon$ seconds, where T_ε has a Geometric distribution with mean $1/\varepsilon$. As a result: $\lim_{\varepsilon\to 0^+}\mathbb{P}[\varepsilon T_\varepsilon > t]=\exp(-t)$, for $t>0$. Thus, as ε gets smaller, the time-separations between consecutive samples from the

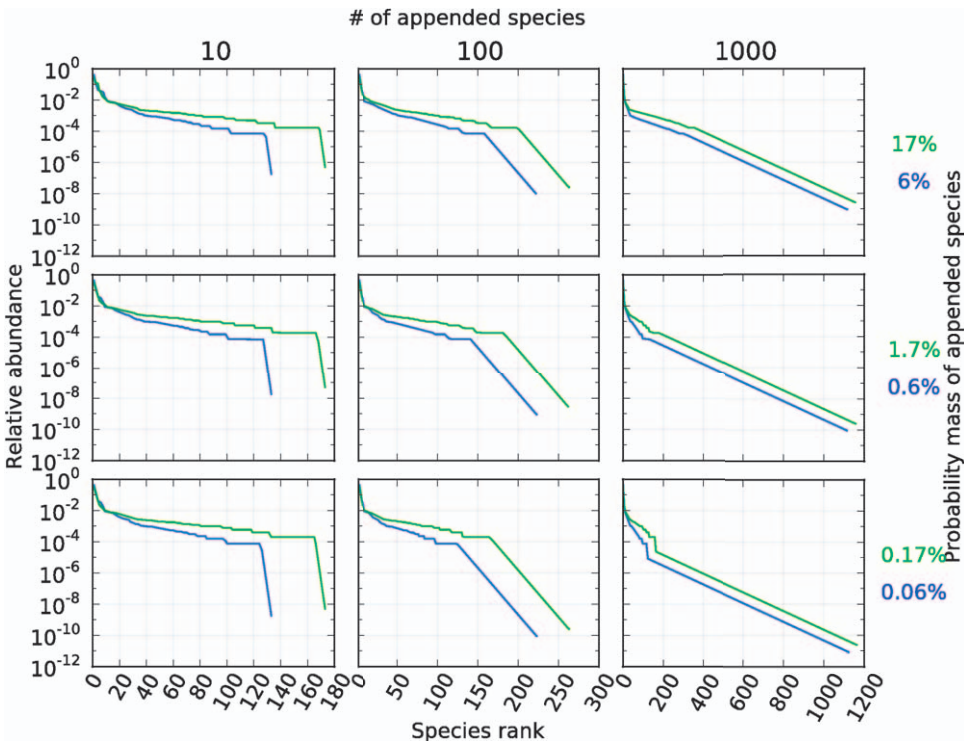

Figure 5. Rank curves associated with the rare biosphere simulation in the human-gut and -hand urn. Rank curves associated with Fig. 6 (green) and Fig. 7 (blue).
doi:10.1371/journal.pone.0021105.g005

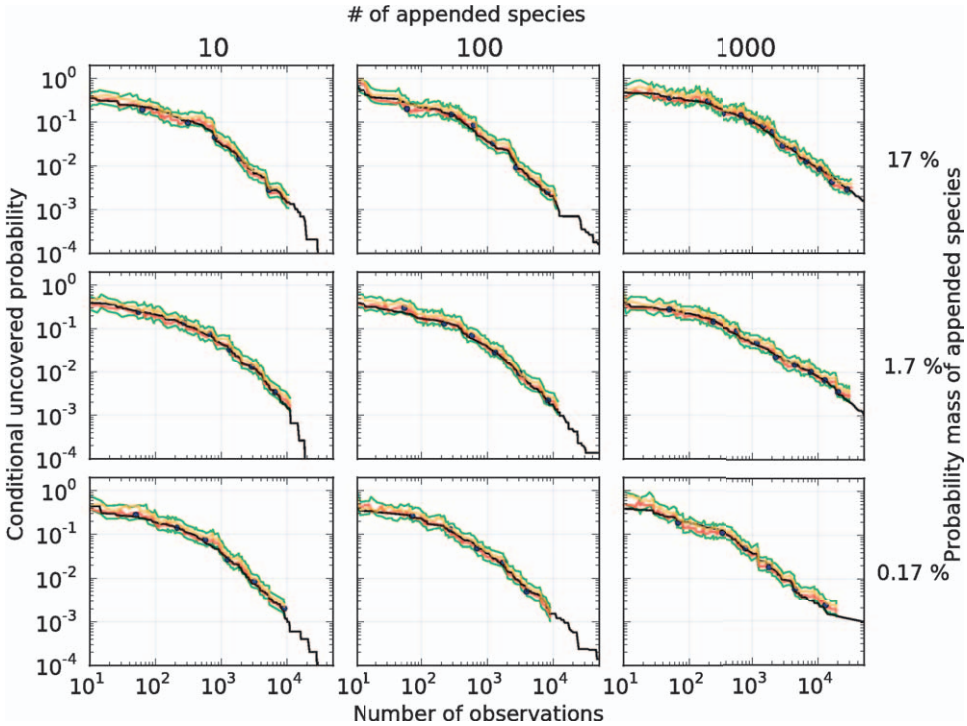

Figure 6. Predictions in the human-hand urn when simulating the rare biosphere. Prediction of the conditional uncovered probability (black) in nine urns associated with a human-hand urn. Point predictions produced by the Embedding algorithm (blue), point predictions produced by the algorithm each time a new species was discovered (red), 95% upper-bound interval (orange), and 95% conservative-upper interval (green). The algorithm used the parameters $(r,f)=(50,2)$. The different urns were devised as follows. For each $i=0.17,0.017,0.0017$ (indexing rows) and $j=10,100,1000$ (indexing columns), a mixture of two urns was considered: an urn with the same distribution as the microbes found in a sample from a human-hand and weighted by the factor $(1-i)$, and an urn consisting of j colors (disjoint from the hand urn), with an exponentially decaying rank curve and weighted by the factor i. See Fig. 5 for the rank curve associated with each urn.
doi:10.1371/journal.pone.0021105.g006

original urn resemble independent Exponential random variables with mean one. The black-urn can therefore be removed from the heuristic altogether by embedding samples from the original urn into a HPP with intensity one over the interval $[0,+\infty)$.

Simulating draws with replacement

To simulate draws with replacement using data already collected from an environment, produce a random permutation of the data. This can be accomplished with low-memory complexity using the discrete inverse transform method to simulate draws–without replacement–from a finite population [31].

Constants associated with optimal prediction intervals

To numerically approximate a pair of constants $0<a<b<\infty$ such that $\int_a^b x^k e^{-x}/k!dx=c$ and $a^k e^{-a}=b^k e^{-b}$, where the integer $k\geq 1$ and the number $0<c<1$ are given constants, introduce the auxiliary variable $t=b/a$, and note that the later condition is fulfilled only when $a=k\cdot\ln(t)/(t-1)$ and $b=t\cdot a$. Due to Newton's method, the sequence $(t_n)_{n\geq 0}$ defined recursively as follows converges to the unique t that satisfies the integrability condition, provided that t_0 is chosen sufficiently close to t:

$$a_n=\frac{k\cdot\ln(t_n)}{t_n-1};$$

$$b_n=t_n\cdot a_n;$$

$$t_{n+1}=t_n\left\{1-\frac{(k-1)!}{a_n^k\cdot e^{-a_n}}\cdot\left(\int_{a_n}^{b_n}\frac{x^k}{k!}e^{-x}dx-c\right)\right\}.$$

Proof of Inequality (3)

First notice that

$$|v_{n,1}-v_{n,r}|\leq\left|\frac{N(1,n+1)}{n+1}-\frac{N(1,n+r)}{n+r}\right| + \sum_{k=2}^{r}\frac{\binom{r-1}{k-1}}{\binom{n+r}{k}}\cdot N(k,n+r). \tag{13}$$

To bound the first term on the right-hand side above, notice that $|N(1,n+1)-N(1,n+r)|\leq(r-1)$. As a result, since $N(1,n+r)\leq(n+r)$, we obtain that:

$$\begin{aligned}&\left|\frac{N(1,n+1)}{n+1}-\frac{N(1,n+r)}{n+r}\right|\\&=\left|\frac{N(1,n+1)-N(1,n+r)}{n+1}+N(1,n+r)\left\{\frac{1}{n+1}-\frac{1}{n+r}\right\}\right|,\\&\leq\frac{r-1}{n-1}+\frac{N(1,n+r)}{n+r}\cdot\frac{r-1}{n+1}\leq\frac{2(r-1)}{n+1}.\end{aligned}\tag{14}$$

On the other hand, to bound the second term on the right-hand side of equation (13), define the quantity $N=\sum_{k=2}^{r}k\cdot N(k,n+r)$ and notice that $N\leq\sum_{k=1}^{n+r}k\cdot N(k,n+r)\leq(n+r)$. Using that a

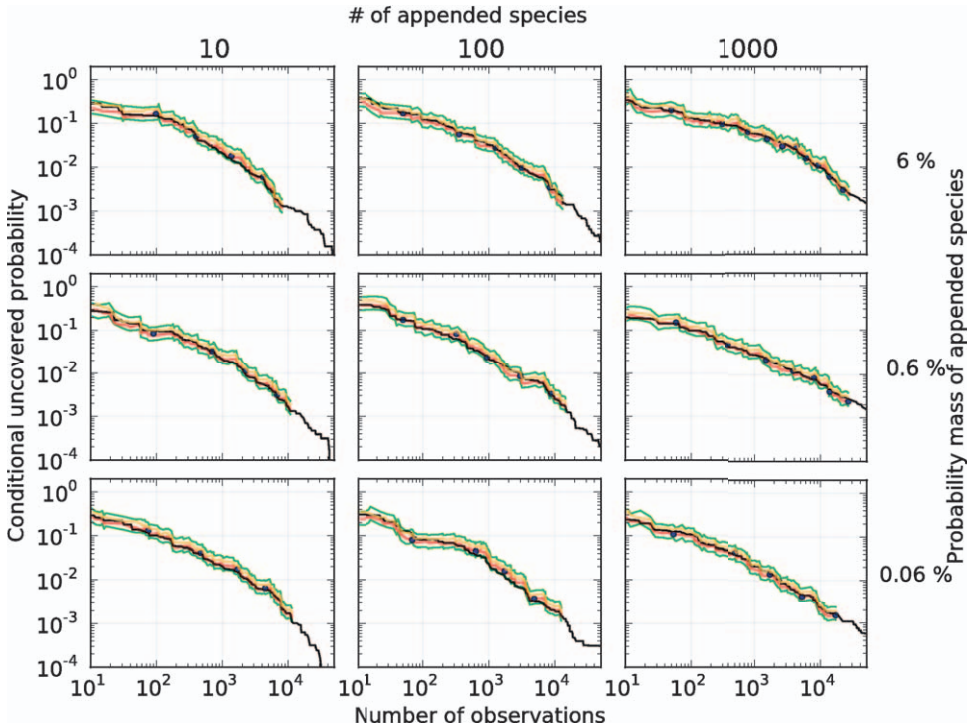

Figure 7. Predictions in the human-gut urn when simulating the rare biosphere. In a sample of size 12,903 from a human-gut, 123 species were discovered. Based on our methods, we estimate that 97 of these species represent $\sim$99.4% of that gut environment; hence, the remaining $\sim$0.6% is composed by at least 26 species. To test our predictions of the conditional uncovered probability (black), we simulated the rare biosphere by adding additional species and hypothesized that our point prediction could be offset by up to one order of magnitude: point predictions produced by the Embedding Algorithm (blue), point predictions produced by the algorithm each time a new species was discovered (red), 95% upper-bound (orange), and 95% conservative-upper interval (green). The predictions used the parameters $(r,f)=(50,2)$. The different urns were devised as follows. For each $i=0.06,0.006,0.0006$ (indexing rows) and $j=10,100,1000$ (indexing columns), a mixture of two urns was considered: an urn with the same distribution as the microbes found in the gut dataset, and weighted by the factor $(1-i)$, and an urn consisting of j colors (disjoint from the gut urn), with an exponentially decaying rank curve and weighted by the factor i. See Fig. 5 for the rank curve associated with each urn.
doi:10.1371/journal.pone.0021105.g007

weighted average is at most the largest of the terms averaged, we obtain that:

$$\begin{aligned}
&\sum_{k=2}^{r}\frac{\binom{r-1}{k-1}}{\binom{n+r}{k}}\cdot N(k,n+r)\\
&=\frac{N}{n+r}\cdot\sum_{k=2}^{r}\frac{(n+r)\binom{r-1}{k-1}}{k\binom{n+r}{k}}\cdot\frac{k\cdot N(k,n+r)}{N},\\
&\leq\max_{2\leq k\leq r}\frac{(n+r)\binom{r-1}{k-1}}{k\binom{n+r}{k}},\\
&=\max_{2\leq k\leq r}\prod_{i=1}^{k-1}\frac{r-k+i}{r-k+i+n}\leq\frac{r-1}{r-1+n},
\end{aligned}\qquad(15)$$

where, for the last inequality, we have used that for each k, the associated product is less or equal to the factor associated with the index $i=(k-1)$. Equation (3) is now a direct consequence of equations (13), (14) and (15).

Proof of Theorem 1

In what follows, f^{-1} denotes the inverse function of f.

Define M to be the set of decreasing partitions of n i.e. vectors of the form $(i_1,\ldots,i_k)$, with $k\geq 1$ and $i_1\geq\cdots\geq i_k\geq 1$ integers, such that $i_1+\ldots+i_k=n$. To each possible sample $(x_1,\ldots,x_n)$, let $g(x_1,\ldots,x_n)$ be the decreasing partition of n associated with the observed ranks in the sample.

Define $p_I=\sum_{i\in I}p_i$, for each set I of colors. Part (i) in the theorem is equivalent to the existence of a function $h:\mathcal{M}\rightarrow[-\infty,\infty]$ such that

$$\mathbb{E}[h(g(X_1,\ldots,X_n))|(X_1,\ldots,X_n)]=f(p_{\{X_1,\ldots,X_n\}}),\qquad(16)$$

with probability one. This is because, in the non-parametric setting, the different colors in the urn carry no intrinsic meaning apart from being different. If there is a certain function h which satisfies condition (16) then $f^{-1}(h((n)))=p_i$, for each color i such that $p_i>0$. In particular, the set $\{j\geq 1\ \textit{such that}\ p_j>0\}$ must be finite. Furthermore, if this set has cardinality l then $p_j=1/l$, for each color j in the set; in particular, $f^{-1}(h((n)))=1/l$. Condition (ii) is therefore necessary for condition (i). Conversely, if condition (ii) is satisfied and the urn is composed by l colors occurring in

equal proportions then the function $h:\mathcal{M}\to[-\infty,\infty]$ defined as $h(i_1,\ldots,i_k)=f(k/l)$ satisfies condition (16).

Proof of Theorem 2

Conditioned on the set I, and the random index i used in Step 3 of the Embedding algorithm, T_r has a Gamma distribution with shape parameter i and scale parameter 1. However, because i has a Negative Binomial distribution, conditioned on I alone, T_r has Gamma distribution with shape parameter r and scale parameter $1/(1-p_I)$. In particular, $(1-p_I)\cdot T_r$ has probability density function $x^{r-1}e^{-x}/(r-1)!$, for $x\geq 0$. From this, parts (i) and (iii) in the theorem are immediate. To show part (ii), notice first that $L_r=(c_r-\ln(T_r))$ is conditionally unbiased for $\ln(1-p_I)$, where

$$c_r=\int_0^\infty \ln(x)\cdot\frac{x^{r-1}e^{-x}}{(r-1)!}\,dx=\frac{1}{r-1}+c_{r-1}.$$

The second identity above is due to an integration by parts argument and only holds for $r\geq 2$. However, since $c_1=-\gamma$, we obtain that $c_r=-\gamma+\sum_{i=1}^{r-1}1/i$, for $r\geq 1$. This shows that L_r is conditionally unbiased for $\ln(1-p_I)$. To complete the proof of the theorem, notice that L_r and $\ln((1-p_I)\cdot T_r)$ have the same variance. In particular, $\mathbb{V}(L_r)=d_r-c_r^2$, where

$$d_r=\int_0^\infty (\ln(x))^2\cdot\frac{x^{r-1}e^{-x}}{(r-1)!}\,dx=\frac{2c_{r-1}}{r-1}+d_{r-1}.$$

The last identity above holds only for $r\geq 2$. Using that $d_1=\gamma^2+\pi^2/6$, we conclude that $d_r=\gamma^2+\pi^2/6+2\sum_{i=1}^{r-1}c_i/i$, for $r\geq 1$. As a result: $\mathbb{V}(L_r)=\pi^2/6-\sum_{i=1}^{r-1}1/i^2$; in particular, since $\sum_{i=1}^{\infty}1/i^2=\pi^2/6$, $\mathbb{V}(L_r)=\sum_{i=r}^{\infty}1/i^2$. The theorem is now a consequence of the following inequalities:

$$\frac{1}{r}=\int_r^\infty\frac{1}{x^2}dx\leq\mathbb{V}(L_r)\leq\int_{r-1}^\infty\frac{1}{x^2}dx=\frac{1}{r-1}.$$

Proof of Equation (8)

Let $z=z_{\alpha/2}$ and assume that $0<z<\sqrt{r-1}$. Observe that:

$$c=\frac{\sqrt{2\pi}(r-1)^{r-1/2}e^{1-r}}{(r-1)!}\cdot\int_{-z}^{z}\frac{e^{-x\sqrt{r-1}}}{\sqrt{2\pi}}\left\{1+\frac{x}{\sqrt{r-1}}\right\}^{r-1}dx.$$

The factor multiplying the previous integral is an increasing function of r; in particular, due to Stirling's formula, it is bounded by 1 from above. Furthermore, from section 6.1.42 in [32], it follows that

$$e^{\frac{-1}{12(r-1)}}\leq\frac{\sqrt{2\pi}(r-1)^{r-1/2}e^{1-r}}{(r-1)!}\leq 1.$$

On the other hand, if one rewrites the integrand of the previous integral in an exponential-logarithmic form and uses that $y-y^2/2+c_{z,r}/(r-1)\leq\ln(1+y)\leq y-y^2/2+y^3/3$, for all $y\geq -z/\sqrt{r-1}$, where

$$c_{z,r}=z\cdot\sqrt{r-1}+\frac{z^2}{2}+(r-1)\cdot\ln\left(1-\frac{z}{\sqrt{r-1}}\right),$$

one sees that

$$e^{c_{z,r}-x^2/2}\leq e^{-x\sqrt{r-1}}\left\{1+\frac{x}{\sqrt{r-1}}\right\}^{r-1}\leq e^{\frac{z^3}{3\sqrt{r-1}}-x^2/2}.$$

All together, these inequalities imply that

$$e^{c_{z,r}-\frac{1}{12(r-1)}}\int_{-z}^{z}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx\leq c\leq e^{\frac{z^3}{3\sqrt{r-1}}}\int_{-z}^{z}\frac{e^{-x^2/2}}{\sqrt{2\pi}}dx,$$

from which the result follows.

Proof of Theorem 3

Due to the Central Limit Theorem, if $c(r)=r-\sqrt{r}\cdot z_{\alpha/4}$ and $b(r)=r+\sqrt{r}\cdot z_{\alpha/4}$ then

$$\lim_{r\to\infty}\int_{c(r)}^{b(r)}\frac{x^{r-1}e^{-x}}{(r-1)!}dx=1-\frac{\alpha}{2}.$$

As a result, for all r sufficiently large, $0\leq b(r)\leq f\cdot c(r)$, and the integral on the left-hand side above is greater than or equal to $(1-\alpha)$. Fix any such r. Since the value of the associated integral may be decreased continuously by increasing the parameter $c(r)$, there is $a(r)$ such that $c(r)\leq a(r)\leq b(r)$ and

$$\int_{a(r)}^{b(r)}\frac{x^{r-1}e^{-x}}{(r-1)!}dx=(1-\alpha).$$

Define $g(t)=\int_t^{f\cdot t}x^{r-1}e^{-x}/(r-1)!dx$, for $t\geq 0$. Since $g(0)=0$ and, because $b(r)\leq a(r)\cdot f$, $g(a(r))\geq(1-\alpha)$, the continuity of $g(\cdot)$ implies that there is $0\leq t\leq a(r)$ such that $g(t)=(1-\alpha)$. Selecting $a=t$ and $b=f\cdot t$ shows the theorem.

Proof of Theorem 4

The proof is based on a coupling argument. First observe that one can define on the same probability space random variables $M,N,E_1,E_2,\ldots$ such that (1) M and N have Negative Binomial distributions with parameters $(r,1-p_I)$, but with M conditioned to be less than or equal to k; (2) $M\leq N$ but $M=N$ when $N\leq k$; and (3) $E_1,E_2,\ldots$ are independent Exponentials with mean 1 and independent of (M,N).

Let A be the event "r balls with colors outside I are observed in the next k draws from the urn". Conditioned on I, we have that $c=\mathbb{P}[a\leq(1-p_I)\cdot T_r\leq b|A]$ and $(1-\alpha)=\mathbb{P}[a\leq(1-p_I)\cdot T_r\leq b]$. As a result:

$$c=\mathbb{P}\left[\frac{a}{1-p_I}\leq\sum_{i=1}^{M}E_i\leq\frac{b}{1-p_I}\right];$$

$$(1-\alpha)=\mathbb{P}\left[\frac{a}{1-p_I}\leq\sum_{i=1}^{N}E_i\leq\frac{b}{1-p_I}\right].$$

Since $\sum_{i=1}^{M}E_i\leq\sum_{i=1}^{N}E_i$, and because $M=N$ when $N\leq k$, we obtain that

$$-\mathbb{P}[N>k]\leq c-(1-\alpha)\leq\mathbb{P}\left[\frac{b}{1-p_I}<\sum_{i=1}^{N}E_i\right].$$

From this, the upper-bound in part (i) and both inequalities in part (ii) follow after noticing that $\sum_{i=1}^{N} E_i$ has a Gamma distribution with shape parameter r and scale parameter $1/(1-p_I)$. To show the lower-bound in (i), we again notice that $\sum_{i=1}^{M} E_i \leq \sum_{i=1}^{N} E_i$. In particular, if $a=0$ then

$$c-(1-\alpha)=\mathbb{P}\left[\sum_{i=1}^{M} E_i \leq \frac{b}{1-p_I}\right]-\mathbb{P}\left[\sum_{i=1}^{N} E_i \leq \frac{b}{1-p_I}\right] \geq 0.$$

Proof of Equations (10) and (11)

Consider random variables X and Y and a random vector Z, defined on a same probability space. Assume that X is square-integrable and conditionally unbiassed for Y given Z i.e. $\mathbb{E}(X|Z)=Y$. Furthermore, assume that $\mathbb{V}(Y)>0$ hence $\mathbb{V}(X)>0$. Because Y is also square-integrable and $\mathbb{E}(X)=\mathbb{E}(Y)$, we obtain that

$$cov(X,Y)=\mathbb{E}((X-\mathbb{E}(Y))\cdot(Y-\mathbb{E}(Y))),$$

$$=\mathbb{E}(\mathbb{E}(X-\mathbb{E}(Y)|Z)\cdot(Y-\mathbb{E}(Y))),$$

$$=\mathbb{E}((Y-\mathbb{E}(Y))^2)=\mathbb{V}(Y).$$

Hence $\rho(X,Y)=\sqrt{\mathbb{V}(Y)/\mathbb{V}(X)}$.

Equation (10) follows by considering $X=(r-1)/T_r$, $Y=U_n$ and $Z=(X_1,\ldots,X_n)$. Similarly, equation (11) follows by considering $X=-\ln(T_r)-\gamma+\sum_{i=1}^{r-1}\frac{1}{i}$ and $Y=\ln(U_n)$.

Proof of Inequality (12)

First note that

$$\rho(U_n,v_{n,r})=\frac{cov(U_n,v_{n,r}-v_{n,1})+cov(U_n,v_{n,1})}{\sqrt{\mathbb{V}(U_n)\cdot\mathbb{V}(v_{n,r})}},$$
$$\leq \frac{\mathbb{E}(\{v_{n,r}-v_{n,1}\}^2)/2+\mathbb{V}(U_n)/2+cov(U_n,v_{n,1})}{\sqrt{\mathbb{V}(U_n)\cdot\mathbb{V}(v_{n,1})}}\cdot\sqrt{\frac{\mathbb{V}(v_{n,1})}{\mathbb{V}(v_{n,r})}}. \quad (17)$$

Now observe that $\mathbb{E}(\{v_{n,r}-v_{n,1}\}^2)=O(r^2/n^2)$ because of inequality (13), which implies that $n\cdot\mathbb{E}(\{v_{n,r}-v_{n,1}\}^2)=o(1)$ because $r\ll\sqrt{n}$. On the other hand, because Robbins' and Starr's estimators are both unbiased for u_n, we have $|\sqrt{\mathbb{V}(v_{n,r})}-\sqrt{\mathbb{V}(v_{n,1})}|\leq\sqrt{\mathbb{E}(\{v_{n,r}-v_{n,1}\}^2)}$. Furthermore, according to the proof of Theorem 2 in [21], $\mathbb{V}(v_{n,1})=\Theta(n^{-1})$, therefore

$$\left|\sqrt{\frac{\mathbb{V}(v_{n,r})}{\mathbb{V}(v_{n,1})}}-1\right|=O\left(\frac{r}{\sqrt{n}}\right).$$

As a result, $\lim_{n\to\infty}\sqrt{\mathbb{V}(v_{n,1})/\mathbb{V}(v_{n,r})}=1$. Inequality (12) is now a direct consequence of inequality (17), and the next identities [21]:

$$\lim_{n\to\infty} n\cdot\mathbb{V}(U_n)=\lambda\cdot e^{-\lambda}-\lambda\cdot(1+\lambda)\cdot e^{-2\lambda};$$

$$\lim_{n\to\infty} n\cdot cov(U_n,v_{n,1})=-\lambda^2\cdot e^{-2\lambda};$$

$$\lim_{n\to\infty} n\cdot\mathbb{V}(v_{n,1})=e^{-\lambda}-(1-\lambda+\lambda^2)\cdot e^{-2\lambda}.$$

Acknowledgments

We would like to thank three anonymous referees for their careful reading of our manuscript and their numerous suggestions, which were incorporated in this final version. The authors are also thankful to R. Knight for contributing to an early version of the code, implementing some of the analyses, and commenting on the manuscript.

Author Contributions

Directed the research project: ML. Developed the new statistical method and accompanying mathematics: ML RG. Implemented the methods: JR. Designed the plots: JR ML. Generated the plots: JR. Wrote the manuscript: ML RG.

References

1. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored \rare biosphere". Proc Natl Acad Sci USA 103: 12115–12120.
2. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. Appl Environ Microbiol 67: 4399–4406.
3. Schloss PD, Handelsman J (2004) Status of the microbial census. Microbiol Mol Biol Rev 68: 686–691.
4. Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, et al. (2006) What is the extent of prokaryotic diversity? Phil Trans R Soc Lond 361: 2023–2037.
5. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. Isme J 1: 283–290.
6. Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. Proc Natl Acad Sci USA 103: 117–122.
7. Quince C, Curtis TP, Sloan WT (2008) The rational exploration of microbial diversity. Isme J 2: 997–1006.
8. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2007) A core gut microbiome in obese and lean twins. Nature 457: 480–484.
9. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, et al. (2010) Forensic identification using skin bacterial communities and/or references within. Proc Natl Acad Sci USA 107: 6477–6481.
10. Magurran AE (2004) Measuring Biological Diversity Oxford - Blackwell.
11. Burnham KP, Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65: 625–633.
12. Chao A (1984) Nonparametric estimation of the number of classes in a population. Scand J Stat 11: 265–270.
13. Chao A (1897) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43: 783–791.
14. Mao CX, Lindsay BG (2007) Estimating the number of classes. Ann Stat 35: 917–930.
15. Bunge J, Fitzpatrick M (1993) Estimating the number of species: A review. J Am Stat Assoc 88: 364–373.
16. Hinsley F, Stripp A (1993) Codebreakers: The Inside Story of Bletchley Park Oxford Univ. Press.
17. Finch SJ, Mendell NR, Thode Jr. HC (1989) Probabilistic measures of adequacy of a numerical search for a global maximum. J Am Stat Assoc 84: 1020–1023.
18. Mao CX (2004) Predicting the conditional probability of discovering a new class. J Am Stat Assoc 99: 1108–1118.
19. Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40: 237–264.
20. Robbins HE (1968) On estimating the total probability of the unobserved outcomes of an experiment. Ann Math Stat 39: 256–257.

21. Starr N (1979) Linear estimation of the probability of discovering a new species. Ann Stat 7: 644–652.
22. Clayton MK, Frees EW (1987) Nonparametric estimation of the probability of discovering a new species. J Am Stat Assoc 82: 305–311.
23. Esty WW (1983) A Normal limit law for a nonparametric estimator of the coverage of a random sample. Ann Statist 11: 905–912.
24. Aldous D (1988) Probability Approximations via the Poisson Clumping Heuristic Springer-Verlag.
25. Mahmoud HM (2000) Sorting: A Distribution Theory Wiley-Interscience.
26. Hwang HK, Janson S (2008) Local limit theorems for finite and infinite urn models. Ann Probab 36: 992–1022.
27. Mao CX, Lindsay BG (2002) A poisson model for the coverage problem with a genomic application. Biometrika 89: 669–681.
28. Durrett R (1999) Essentials of stochastic processes Springer Texts in Statistics.
29. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, et al. (2009) The e_ect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 1: 6ra14.
30. Fierer N, Hamady M, Lauber CL, Knight R (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci USA 105: 17994–17999.
31. Ross SM (2002) Simulation Academic Press, third edition.
32. Abramowitz M, Stegun IA (1964) Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover, ninth Dover printing, tenth GPO printing edition.