HHS Public Access

Author manuscript

Quant Biol. Author manuscript; available in PMC 2016 May 30.

Published in final edited form as:

Quant Biol. 2015 September ; 3(3): 135–144. doi:10.1007/s40484-015-0049-7.

Applications of species accumulation curves in large-scale biological data analysis

Chao Deng, **Timothy Daley**, and **Andrew D Smith***

Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

Abstract

The species accumulation curve, or collector's curve, of a population gives the expected number of observed species or distinct classes as a function of sampling effort. Species accumulation curves allow researchers to assess and compare diversity across populations or to evaluate the benefits of additional sampling. Traditional applications have focused on ecological populations but emerging large-scale applications, for example in DNA sequencing, are orders of magnitude larger and present new challenges. We developed a method to estimate accumulation curves for predicting the complexity of DNA sequencing libraries. This method uses rational function approximations to a classical non-parametric empirical Bayes estimator due to Good and Toulmin [Biometrika, 1956, 43, 45–63]. Here we demonstrate how the same approach can be highly effective in other large-scale applications involving biological data sets. These include estimating microbial species richness, immune repertoire size, and *k*-mer diversity for genome assembly applications. We show how the method can be modified to address populations containing an effectively infinite number of species where saturation cannot practically be attained. We also introduce a flexible suite of tools implemented as an R package that make these methods broadly accessible.

Keywords

species accumulation curve; accumulation region; rational function approximation; immune repertoire; microbiome diversity; species richness

INTRODUCTION

In ecology the species richness, the total number of species or distinct classes, is one of the simplest metrics for understanding the diversity of a population [1]. However, unbiased estimation of species richness based on surveys is often extremely difficult or even unattainable because no matter how many species have been observed there may exist an arbitrary number of undetected rare species in the population [2]. An alternative is to study

*Correspondence: ; Email: andrewds@usc.edu

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at DOI 10.1007/s40484-015-0049-7.

The authors Chao Deng, Timothy Daley and Andrew D Smith declare they have no conflict of interest.

COMPLIANCE WITH ETHICS GUIDELINES

This article does not contain any studies with human or animal subjects performed by any of the authors.

the expected number of unique species as a function of the size of the survey, defined as the species accumulation curve [3]. Though commonly applied in ecological studies, accumulation curves have been applied to many other fields, such as linguistics [4], genetics [5], metagenomics [6], and immune repertoire [7].

Consider the following problems as motivation. Researchers want to compare the diversities of several populations. In each population, individuals are sampled and their corresponding species identified but the total number of individuals sampled can vary across populations. In these sampling experiments, often called “capture-recapture” experiments, the raw data are the sample, or “capture”, counts associated with each species. Since in general the number of species observed increases with the number of captured individuals and the size of the survey, direct comparison between surveys can bias the result. With species accumulation curves, one can make fair comparisons of the expected number of species for a fixed number of individuals captured [8]. Another problem is to evaluate the effectiveness of a survey and decide whether or not to continue the project. A typical question might be: given capture profiles in previous surveys, if another survey is conducted from the same population, how many new species can researchers expect to sample? Accurate predictions can help scientists make better decisions and allocate resources more appropriately.

Various models have been proposed to address these problems and are well discussed by Bunge and Fitzpatrick [2] and Colwell et al. [9].We assume a general compound Poisson mixture model for the sampling process. In particular, for each species in the underlying population the number of individuals observed in a survey follows a Poisson distribution with the Poisson rate λ generated from a latent distribution $G(\lambda)$.

Many approaches rely upon inferring the latent distribution $G(\lambda)$. These are further divided into parametric and non-parametric approaches. The former assumes $G(\lambda)$ takes a particular form. For example, Fisher et al. adopted a Gamma distribution; Bulmer used a log-normal; and Burrell and Fenton applied a generalized inverse Gaussian [10–12]. A problem with this approach is that it is difficult to assess a parametric form given the data. Several parametric models may fit the data equally well, but their extrapolation behaviors can be quite different [13].

On the other hand, most non-parametric approaches approximate $G(\lambda)$ with a discrete distribution [14–16]. It can be shown that due to the discrete nature of the observed data, the likelihood achieves its maximum at a discrete distribution [17]. However, non-parametric approaches tend to underestimate the number of unique species due to inadequate sampling efforts and skewness of the abundance curve [18].

Good and Toulmin addressed the problem of predicting accumulation curves under a general multinomial model [19]. They derived an unbiased non-parametric empirical Bayes estimator for the gain in new species as a function of the relative increase in survey size. This estimator, which we call the Good-Toulmin estimator, takes the form of an alternating power series and avoids direct inference of the latent distribution. This method was later extended by Efron and Thisted to a general compound Poisson model [4]. The Good-Toulmin estimator can very accurately predict the number of new species gained when the

survey is increased to up to twice the initial size. Unfortunately due to its alternating form, the power series diverges when extrapolating beyond the twice the size of the initial survey. This short range of extrapolation makes the estimator useless in most applications. To partially overcome this difficulty, Good and Toulmin suggested using Euler's series transformation. This increases the practical radius of convergence but the range of extrapolation is still very limited [4,20].

Rational function approximations were proposed by Daley and Smith to address the divergence problem observed in the Good-Toulmin power series [21]. The constructed rational function approximations have two critical properties: (i) the local behavior of the rational function approximation is close to the Good-Toulmin power series in a sense that they have the same Taylor expansion centered at the size of the observed experiment up to a fixed degree; and (ii) global stability by choice of the degree of the rational function approximation. Previous applications to DNA sequencing libraries showed that rational function approximations can accurately extrapolate up to one hundred times the size of the initial survey [21,22], significantly further than previous methods.

Our goal here is to present additional applications of accumulation curves in the analysis of large-scale biological data sets. These applications will serve to introduce researchers to the general class of problems that can be modeled as sampling experiments aiming to make inferences about heterogeneity within poorly-understood populations. At the same time, these applications serve to demonstrate the broad applicability of our approach via rational function approximation to the Good-Toulmin estimator. We have developed an R package that implements all the functionality required to conduct the analysis presented here; all figures and results we present have been generated using this package, with all steps included in Supplementary Materials.

The rest of this paper is organized as follows. First we review the concepts associated with the Good-Toulmin estimator and use of rational function approximations for computing it. We then apply this method to three biological data analysis problems arising in different contexts: investigating bacterial species diversity of a metagenomic sample, estimating the size of immune repertoire through T cell receptor (TCR) diversity, and examining *k*-mer diversity of next generation sequencing reads for genome assembly problems.

RESULTS

The general compound Poisson model

We assume a general compound Poisson sampling model, and follow the notation of Wang [23]. In a survey or sampling experiment, observations are usually recorded as x_i, $i = 1, 2, \ldots, D$, indicating $x_i > 0$ individuals observed from species i for D total observed species. The data can be further summarized into a vector of frequencies $n = (n_1, n_2, \ldots)$, where

$$n_j = \sum_{i=1}^{D} I\{x_i = j\}$$

represents the number of species with exactly j individuals captured for each species. The number n_0 of unobserved species is unknown and is non-identifiable in the general compound Poisson model [24].

One well known application of accumulation curves under a general compound Poisson model is Shakespeare's vocabulary [4]. The word frequencies in Shakespeare's known works, taken from the Ref. [4], are summarized in Table 1. A total of $D = \sum_j n_j = 31534$ unique words are observed. Among those words, 14,376 appear once, 4,343 words appear twice, and so on.

Based on the frequency data, we can estimate the expected number of species (which are the unique words in the aforementioned application), as a function of the total number of individuals observed. The Good-Toulmin power series is an unbiased non-parametric empirical Bayes estimator for the expected number of new species that will be observed with further sampling under a multinomial or compound Poisson model. This power series is given by

$$\hat{\Delta}(t)=\sum_{j=1}^{\infty}(-1)^{j+1}(t-1)^j n_j, \quad (1)$$

where t is the relative size of the survey and the initial observed experiment corresponds to $t = 1$.

The Good-Toulmin power series has a radius of convergence of 1 centered at $t = 1$. As a consequence, the estimator performs extremely well up to a doubling of the size of the experiment, the range $1 < t \leq 2$, but commonly diverges for $t > 2$ [19]. The application of rational function approximations eliminates the divergence problem of the Good-Toulmin power series (1) while retaining the desirable local properties of the power series.

In most experiments, the populations under investigation contain a finite number of classes. To ensure the asymptotic behavior of the rational function approximation matches the asymptotic behavior of the population, we choose the rational function approximation to be constant in the limit. That is, for some integer M,

$$\hat{\Delta}(t) \approx (t-1)\frac{p_0+p_1(t-1)+\cdots+p_{M-1}(t-1)^{M-1}}{1+q_1(t-1)+\cdots+q_M(t-1)^M}. \quad (2)$$

This ensures the stability of long-range extrapolations of the species accumulation curve. The coefficients of the rational function approximation are chosen so that the power series of the rational function approximation matches the observed power series and the local behavior will be similar [25].

We implement the rational function approximations through continued fractions. Continued fraction approximations can be shown to be equivalent to rational function approximations as shown in Equation (2) but have several advantages. These include asymptotically faster and more stable computation of the coefficients (quadratic versus cubic), stable evaluation of the approximation for large t with Euler's recursion, and the ability to easily adjust the

degree of the rational function. For example, a continued fraction approximation to $\hat{\Delta}(t)$ with four coefficients has the form:

$$n_1(t-1) - n_2(t-1)^2 + n_3(t-1)^3 - n_4(t-1)^4 \approx \frac{a_0(t-1)}{1+\frac{a_1(t-1)}{1+\frac{a_2(t-1)}{1+a_3(t-1)}}}. \quad (3)$$

To fit a rational function approximation we search the space of continued fraction approximations for one lacking apparent instabilities, known as defects, that arise as a consequence of using stochastic estimates of the coefficients instead of the true coefficients of the Good-Toulmin power series [26]. We test the curve for two critical properties, that it is concave and nondecreasing. If a defect is found, we test a different order approximation, up to the maximum degree specified. Further details can be found in the Ref. [27].

For interpolation of species accumulation curves, we explicitly calculate the expected number of species in a subsample of n individuals, denoted D_n. This expectation is equal to

$$E(D_n) = D_N - \binom{N}{n}^{-1} \sum_{i=1}^{D_N} \binom{N-n_i}{n}, \quad (4)$$

where N is the total number of observed individuals in the original sample [28].

For simplicity and brevity we shall abbreviate our approach of taking rational function approximations to the Good-Toulmin power series as RFA-GT. For flexible use of this approach, we have implemented a set of functions in an R package, called preseqR and available through CRAN (http://cran.r-project.org/web/packages/preseqR/index.html). All results presented here were obtained using functions in preseqR (which contains broader functionality than the RFA-GT). For all applications presented below, we provide code and data in the Supplementary Materials to allow for exact reproduction of all results.

Revisiting Shakespeare's vocabulary

We revisited a famous application of capture-recapture models constructed by Efron and Thisted [4]. They investigated the vocabulary of Shakespeare using a bag of words sampling model to estimate the number of words that Shakespeare knew but were not used in his known works. The "species" in this application are unique words in Shakespeare's vocabulary and the accumulation curve is the number of unique words as a function of total words in Shakespeare's works.

To evaluate the performance of our method against other existing methods, we sampled approximately 5% and 10% of the total words from Shakespeare's known works as initial surveys (44,000 and 88,000 words, respectively). The rarefaction curve from all of Shakespeare's known works is considered as the gold standard for comparison. The scaled error, defined by the difference between the estimated value and the true value divided by the true value, measures the accuracy of the predictions. We compared the performance of RFA-GT implemented with preseqR to another state of the art method, the R package iNEXT [29], whose extrapolation method is described in the Ref. [9]. All calculations were made on a 2.6 GHz Intel Core i5 with 8 Gb of RAM. It took about 34 s and 16 s for preseqR

to construct the accumulation curve for the 5% and 10% samples as initial surveys. More running time was spent on the 5% sample because over twice as many iterations was required to generate 100 bootstraps without apparent defects as the 10% sample (264 versus 121). iNEXT took around 15 s and 32 s for the 5% and 10% samples, respectively.

Using the 5% sample as the initial survey, both methods accurately estimated the expected number of unique words when extrapolating to two or three times the size of the initial survey. However, as the extrapolation went further, RFA-GT gave much better predictions than iNEXT (Figure 1A). The estimated number of unique words for each of the methods and the scaled errors are provided in Table 2. For both methods, the scaled error increases as the extrapolation increases. However, RFA-GT had less than 10% scaled error extrapolating to up to 10 times the initial experiment while iNEXT had 36.7% scaled error. When extrapolating to 20 times, the scaled error for RFA-GT was 19.7% compared to 53.9% for iNEXT, a difference of over 10,000 words.

Increased sample size gives more information on the population so that we expect predictions to be better with increased sample size, even to the same relative extrapolation. As expected, the predictions from both methods improved when the size of the initial survey doubled to 10%. RFA-GT underestimated the vocabulary by 2,300 words while iNEXT underestimated by over 10,000 when extrapolating to 10 times the initial survey, a difference of nearly 30% of the total vocabulary (Figure 1B and Table 3).

To evaluate the consistent performance of the estimators we repeated each of the above experiments for 100 independent samples. For the 5% samples the mean absolute deviation (MAD) from the true vocabulary size over the 100 samples were 4,900 and 16,900 words for RFA-GT and iNEXT, respectively. For the 10% samples the MAD was 3,400 and 11,600 words. This shows the consistently better performance of RFA-GT that may be due to a number of reasons. The Good-Toulmin power series uses the full amount of information contained in the observed experiment, without smoothing. Rational function approximations can be considered to be the optimal smoothing of the power series that satisfies the constraint that the estimated accumulation curve is asymptotically constant [25].

Based on the success of RFA-GT in predicting Shakespeare's vocabulary, we used the entirety of Shakespeare's known works as an initial survey to estimate the expected number of total distinct words we would observe if new works of Shakespeare were to be discovered. Shakespeare's known works contain 31,534 unique words. If the size of Shakespeare's works were to double, we estimated a total of 43,038 unique words observed (Figure 1C). Thus if an additional volume of works by Shakespeare were to be discovered, equal in size to his currently known works, we can expect an additional 43038 – 31534 = 11504 new words. The result is quite close to the reported 11,430 words in the Ref. [4], from the equation (2.5), since the behavior of rational function approximation is close to the Good-Toulmin power series when t is small.

As Figure 1 shows and we have previously observed [21,27], RFA-GT tends to produce accurate but conservative estimates. Thus we can use the extrapolation to ten times of the totality of Shakespeare's known works as a reasonable lower bound for the total vocabulary

of Shakespeare. RFA-GT estimated a total of 75,722 unique words, 9,188 more unique words than the lower bound estimated by Efron and Thisted [4].

Species richness in a microbiome

Capture-recapture models are a common statistical model for estimating microbial species richness [6,30]. We applied preseqR to the problem of estimating accumulation curves of annotated species in a metagenomic sequencing experiment. We examined species abundance data collected and calculated by Yatsunenko et al. [31], downloaded from MG-RAST with ID 4461119.3 [32]. The data contains 1,712 unique annotated species with a total sampled abundance of 156,608. Though the abundance of annotated species underestimates the total diversity of the sample, due to overwhelming presence of unannotated species in the microbial universe, it can be used as a proxy metric or to compare samples.

To test RFA-GT on microbial species diversity we subsampled a total abundance of 7,830 without replacement as an initial experiment, approximately 5% of the full experiment. We compare our estimated curve against the true curve provided by MG-RAST. The true curve is given in terms of the number of sequences. We scale this curve assuming that the observed abundance of genomic material arising from annotated species is proportional to the number of sequenced reads. Note that if our assumption is incorrect then we should see higher error in our estimates.

When we compared the predicted results with the true species abundance curve, we saw that the RFA-GT accurately estimated the species accumulation curve (Figure 2A). RFA-GT predicted 1,631 unique annotated species if a total abundance of 156,608 is sampled compared to the observed value of 1,712 unique annotated species. This is a difference of less than 5% relative to the total number of unique annotated species.

We applied the same methodology for obtaining a lower bound as we did in estimating the size of Shakespeare's vocabulary (Section 3). We used the full experiment to predict the total number of annotated species in the experiment. The predicted curve asymptotes relatively quickly, indicating that the observed experiment is nearly saturated. The observed experiment has 1,712 annotated species. A ten fold increase in the experiment is expected to only yield an additional 423 species, for a total of 2,135 annotated species. This is quite close to the estimated saturation point of annotated species of approximately 2,230 (Figure 2B). This indicates that the observed experiment is already quite saturated and significant resources will need to be expended to observe additional annotated species.

Age-related decrease in TCR repertoire

We applied our method to investigate age-related decreases in T cell receptor (TCR) diversity. The data sets are profiles of TCR β repertoires in 39 healthy donors aged 6–90 years (y) from Britanova et al. [33]. For each donor, the data is summarized as frequencies of TCR β CDR3 clonotypes. Species in this study are TCR β CDR3 clonotypes and the accumulation curve is the expected number of unique TCR β CDR3 clonotypes as a function of TCR β cDNA molecules sequenced.

We constructed TCR β CDR3 clonotype accumulation curves for each donor. These curves are classified into four groups based on the ages of donors: group 1 is composed of the youngest donors from 6–25 y; group 2 is middle-aged with donors from 34–43 y; group 3 contains donors from 61–66 y; and group 4 are the eldest donors from 71 to 90 y [33], (Table 1). For each group of donors we define an accumulation region as the interval formed by the 30% and 70% quantiles for each age group.

As illustrated in Figure 3A and B, it was clear that the diversity of TCR β CDR3 clonotypes decreased with the age of the group. Groups 1 and 2 are distinguished from each other and the other two groups through their accumulation regions. On the other hand, the median accumulation curves of group 3 and group 4 were almost identical and the ranges overlap, consistent with the results of Britanova et al. [33].

Another interesting feature is that the width of the accumulation region reflects the variation of the diversity of TCR β CDR3 clonotypes among donors in a group. The interpolated accumulation regions widths have no appreciable difference among groups (Figure 3A). However, using the prediction results from preseqR, the width of the accumulation region in group 1 is expected to be much larger than other groups if the experiment were to be continued (Figure 3B). In the observed experiment group 1 has a similar variance in diversity to groups 3 and 4 but when extrapolated out to 20 million total TCR β cDNA molecules, the estimated diversity of group 1 has nearly five times the variance of either groups 3 or 4 and over twice the variance of group 2. Most of this variability is seen to arise from subjects who are 16 years old or younger. One possible interpretation, in line with the observations of the Ref. [34], is that this may indicate high variability in the immune repertoire and presence of a large population of rare TCR β clonotypes in youth, prior to selection due to exposure to pathogens. For conclusive results, a larger sample size of youth immune repertoire might be required.

We tested the performance of RFA-GT by combining all 39 data sets. One million cDNA molecules were sampled without replacement from the combined data set as an initial experiment. Extrapolating out to 20 million total molecules, we predicted 8.97 million unique clonotypes compared to an expected value of 9.73 million. This represents a relative error of less than 10% when predicting out 20 fold of the initial experiment. When we used RFA-GT to predict the full experiment of 38.7 million sequenced molecules, we estimated 13.72 million distinct clonotypes. This is 17.5% lower than the observed diversity of 16.7 million. Although the predictions are less accurate when increasing the range of extrapolation, they tend to be conservative (Figure 3C). It is worth noting that conservative prediction is especially useful when attempting to design an efficient experiment to cover a minimum proportion of target population.

k-mer diversity in genome assembly applications

Many state of the art genome assembly algorithms leverage De Bruijn graphs constructed from *k*-mers that are extracted from the sequenced reads [35]. In order to effectively use these graphs, the assembly algorithms require a sufficient fraction of the *k*-mers from the underlying genome [36]. Clearly deeper sequencing will ensure that more *k*-mers have been sampled but the rate at which deeper sequencing reveals more *k*-mers and the reliability of

such *k*-mers is unknown. Here we show how species accumulation curves can provide information about the diversity of sequenced *k*-mers as a function of total bases sequenced.

We selected a data set from the Assemblathon 2 [37], whole genome sequencing of a male budgerigar (also known as the common pet parakeet or *Melopsittacus undulatus*), Sequenced Read Archive accession number ERX218679. This experiment contains approximately 161 million read pairs (150 bp in length) for a total of approximately 48 billion sequenced bases.

We subsampled two experiments from the sequenced reads, randomly downsampling approximately 1% and 10%of the reads from the full experiment. This resulted in 3.23 million and 32.26 million reads, respectively. We examined *k*-mers for $k = 31$ since it is the default setting for the widely used assembly algorithm Velvet [38]. We counted the *k*-mer occurrences using Jellyfish[39] and used the 31-mer counts from the subsampled experiments to extrapolate the distinct 31-mers as a function of total sequenced 31-mers (or equivalently, total bases sequenced).

For the 1% downsampled experiment, the estimated number of distinct 31-mers for the full experiment is 4.24 billion compared to the 5.66 billion observed. On the other hand, if the extrapolation is limited to 10% of the total experiment, the estimated value by RFA-GT is 2.08 billion compared to an expected value of 2.07 billion, a relative difference of less than 1%.

For the 10% downsampled experiment, the estimated number of distinct 31-mers in the full experiment is 5.3 billion showing a slight decrease in accuracy with the increase in sample size when extrapolating to the same relative size. We should expect the accuracy of the estimates to the same relative extrapolation size to increase with increasing sample size based upon the extra information contained in the larger experiment. Instead we observed the opposite (Figure 4A). RFA-GT closely approximated the curve when the number of total 31-mers is relatively small, but we fail to capture the behavior of the tail of the curve.

We noted that the curve of distinct 31-mers plotted against the total number of observed 31-mers appears to be linear in the limit (Figure 4). This can be explained by the presence of random errors in the sequenced reads. The number of distinct 31-mers in the budgerigar genome is bounded by size of the genome, approximately 1.2 gigabases. Indeed *k*-mers in a genome are far from uniformly distributed [40], resulting in far fewer *k*-mers than is theoretically possible, which is bounded above by the genome size. On the other hand, random errors in the sequencing process can theoretically produce any 31-mer for a total of approximately 4.6×10^{18} possibilities, many orders of magnitude larger than current sequencing experiments. This leads us to hypothesize that there is a large tail in the population of observable 31-mers, mostly due to sequencing error. We reasoned that using a different order of rational function approximation may increase the accuracy.

In the applications we presented in earlier sections, the population was known to be finite. Consequently we chose to use a rational function approximations that were constant in the limit. However, we can also choose a rational function approximation that is linear in the limit, for example:

$$\hat{\Delta}(t) \approx (t-1)\frac{p_0+p_1(t-1)+\cdots+p_M(t-1)^M}{1+q_1(t-1)+\cdots+q_M(t-1)^M}.$$

In the context of counting *k*-mers in a sequencing data set, using a linear limit improved the accuracy of the estimates dramatically. Figure 4 shows this alteration predicted 5.55 billion 31-mers in the full experiment for the 10% downsampled experiment and 5.22 billion for the 1% downsampled experiment when extrapolating to the full experiment.

The asymptotically linear extrapolation may be of interest for populations that are infinite or have extremely long tails, with a vast number of species that have extremely small abundances. In such cases the asymptotic linear slope will be related to rate of discovery of the extremely rare species. In the *k*-mer counting application, the eventual linear behavior is driven by sequencing errors generating a practically limitless supply of *k*-mers that could be sequenced.

DISCUSSION

As high-throughput technologies improve researchers will be increasingly faced with the difficult problem of making inferences on unknown and highly variable populations. When the properties of the population are unknown, capture-recapture models may be appropriate. Here we investigated three applications of capture-recapture models to data arising from next-generation sequencing experiments along with the classical application of estimating the size of Shakespeare's vocabulary. These applications demonstrate the breadth of data analysis contexts in which species accumulation curves, and capture-recapture perspectives more generally, can help to understand the underlying populations from which data have been sampled.

Large-scale applications present new challenges to traditional capture-recapture statistics, particularly since the scale of the data is orders of magnitude larger than classical ecological capture-recapture experiments. Algorithms are required that are both scalable and able to accommodate arbitrary heterogeneity to ensure accurate inference. The non-parametric empirical Bayes estimator, Equation (1), is ideal to accommodate unknown heterogeneity. This estimator has been applied directly to estimate both microbial species richness and TCR diversity [41,42]. In both cases the extrapolation region was limited to a two-fold increase in the size of the existing data set. We have demonstrated the applicability of RFA-GT to both of these situations and that this approach does not suffer from a tightly constrained range of extrapolation. In large-scale applications, such a tightly constrained range of extrapolation has the potential to impact conclusions drawn about the sampled populations, making approaches like ours extremely valuable.

Among the applications we have surveyed, the estimation of *k*-mer diversity has an interesting property: sequencing errors can, in theory, produce any possible *k*-mer. The underlying population is therefore practically infinite. In finite populations, large extrapolations of the species accumulation curve can be considered as a conservative estimate of the "species richness" [43]. But when the underlying population is infinite, the

concept of species richness becomes meaningless, and different approaches are required to understand heterogeneity in the population. By modifying the form of rational functions used to approximate the Good-Toulmin power series, we showed how our approach can model accumulation curves that are linear in the limit.

Despite the flexibility of the RFA-GT approach, and its accuracy for large-scale extrapolation, the approach depends on having sufficient initial sampling to fit the rational function approximation. In the case of finite populations, RFA-GT requires the first four frequency counts (n_1, n_2, n_3, n_4) to be positive. This is because an even number of terms is needed to ensure numerical stability of the estimation algorithm, and two terms is not sufficient to properly account for heterogeneity in the population. In the infinite population case, RFA-GT requires the first five frequency counts to be positive due to the increased difficulty in predicting the species accumulation curve in such populations. So although RFA-GT solves challenges associated with large-scale applications, it also depends on properties of those data sets that present difficulties for other methods. Beyond these requirements on the low-order counts, the procedures we use at present to fit rational functions from counts histograms are intricate and include many steps with the potential for introducing error. Despite the high accuracy we typically observe from our method, several of the numerical procedures it requires are candidates for novel algorithm development to further improve stability.

For ease of use we have created an R package, preseqR, to allow researchers easy and convenient access to the RFA-GT method. preseqR is available through CRAN at: http://cran.r-project.org/web/packages/preseqR. The implementation of preseqR includes several core routines in the form of R extensions, which were written originally in C++ for efficiency. All of the analyses performed in this manuscript with preseqR and we have made all code available as part of the supplementary materials as a guide for researchers.

Biological and biomedical science continues to push towards examining increasingly precise hypotheses using large-scale data production. Although estimating heterogeneity in underlying molecular population is rarely the goal, some understanding of the underlying population may be essential to accurate interpretation of analysis results. Classical capture-recapture statistics has frequently addressed questions analogous to those we now face concerning heterogeneity in molecular populations, and represents a robust body of statistical methodology that warrants broader adoption in large-scale data analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was funded by NIH grant R01 HG 007650. We would like to thank the Smith's lab members, Michael Waterman, Simon Tavaré, and Peter Calabrese for their helpful comments.

REFERENCES

1. Magurran, AE. Ecological Diversity and Its Measurement. Vol. 168. Princeton: Princeton University Press; 1988.
2. Bunge J, Fitzpatrick M. Estimating the number of species: A review. J. Am. Stat. Assoc. 1993; 88:364–373.
3. Colwell RK, Mao CX, Chang J. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology. 2004; 85:2717–2727.
4. Efron B, Thisted R. Estimating the number of unseen species: How many words did Shakespeare know? Biometrika. 1976; 63:435–447.
5. Ionita-Laza I, Lange C, Laird NM. Estimating the number of unseen variants in the human genome. Proc. Natl. Acad. Sci. USA. 2009; 106:5008–5013. [PubMed: 19276111]
6. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: Statistical approaches to estimating microbial diversity. Appl. Environ. Microbiol. 2001; 67:4399–4406. [PubMed: 11571135]
7. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, Darko S, Kroll JS, Douek DC, Price DA, Bangham CR, et al. Quantification of HTLV-1 clonality and TCR diversity. PLoS Comput. Biol. 2014; 10:e1003646. [PubMed: 24945836]
8. Gotelli NJ, Colwell RK. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 2001; 4:379–391.
9. Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, Longino JT. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. 2012; 5:3–21.
10. Fisher RA, Corbet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. 1943; 12:42–58.
11. Bulmer M. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics. 1974; 30:101–110.
12. Burrell QL, Fenton MR. Yes, the GIGP really does work – and is workable! J. Am. Soc. Inf. Sci. 1993; 44:61–69.
13. Engen, S. Stochastic Abundance Models. London: Chapman and Hall; 1978.
14. Norris JL, Pollock KH. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. Environ. Ecol. Stat. 1998; 5:391–402.
15. Wang J-PZ, Lindsay BG. A penalized nonparametric maximum likelihood approach to species richness estimation. J. Am. Stat. Assoc. 2005; 100:942–959.
16. Mao CX, Colwell RK, Chang J. Estimating the species accumulation curve using mixtures. Biometrics. 2005; 61:433–441. [PubMed: 16011689]
17. Lindsay BG. The geometry of mixture likelihoods: A general theory. Ann. Stat. 1983; 11:86–94.
18. Wang J-P. Estimating species richness by a Poisson-compound Gamma model. Biometrika. 2010; 97:727–740. [PubMed: 22822253]
19. Good I, Toulmin G. The number of new species, and the increase in population coverage, when a sample is increased. Biometrika. 1956; 43:45–63.
20. Keating KA, Quinn JF, Ivie MA, Ivie LL. Estimating the effectiveness of further sampling in species inventories. Ecol. Appl. 1998; 8:1239–1249.
21. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. Nat. Methods. 2013; 10:325–327. [PubMed: 23435259]
22. Daley T, Smith AD. Modeling genome coverage in single-cell sequencing. Bioinformatics. 2014; 30:3159–3165. [PubMed: 25107873]
23. Wang J-P. SPECIES: An R package for species richness estimation. J. Stat. Softw. 2011; 40:1–15.
24. Mao CX, Lindsay BG. Estimating the number of classes. Ann. Stat. 2007; 35:917–930.
25. Baker, G.; Graves-Morris, P. Padé Approximants (Encyclopedia of Mathematics and its Applications). 2nd. London: Cambridge University Press; 1996.
26. Baker GA Jr. Defects and the convergence of Padé approximants. Acta Appl. Math. 2000; 61:37–52.

27. Daley, TP. Ph.D. thesis. University of Southern California; 2014. Non-Parametric Models for Large Capture-Recapture Experiments with Applications to DNA Sequencing.
28. Heck KL Jr, van Belle G, Simberloff D. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology. 1975; 56:1459–1461.
29. Hsieh TC, Ma KH, Chao A. iNEXT online: interpolation and extrapolation [software]. 2013 http://chao.stat.nthu.edu.tw/blog/software-download/inext-online/.
30. Bunge J, Willis A, Walsh F. Estimating the number of species in microbial diversity studies. Annu. Rev. Stat. Appl. 2014; 1:427–445.
31. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al. Human gut microbiome viewed across age and geography. Nature. 2012; 486:222–227. [PubMed: 22699611]
32. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9:386. [PubMed: 18803844]
33. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, Bolotin DA, Lukyanov S, Bogdanova EA, Mamedov IZ, et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J. Immunol. 2014; 192:2689–2698. [PubMed: 24510963]
34. Wedderburn L, Patel A, Varsani H, Woo P. The developing human immune system: T-cell receptor repertoire of children and young adults shows a wide discrepancy in the frequency of persistent oligoclonal T-cell expansions. Immunology. 2001; 102:301–309. [PubMed: 11298828]
35. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA. 2001; 98:9748–9753. [PubMed: 11504945]
36. Compeau PE, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. 2011; 29:987–991. [PubMed: 22068540]
37. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. Gigascience. 2013; 2:1–31. [PubMed: 23587291]
38. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18:821–829. [PubMed: 18349386]
39. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011; 27:764–770. [PubMed: 21217122]
40. Ren J, Song K, Deng M, Reinert G, Cannon CH, Sun F. Inference of markovian properties of molecular sequences from NGS data and applications to comparative genomics. Bioinformatics. 2015
41. Kroes I, Lepp PW, Relman DA. Bacterial diversity within the human subgingival crevice. Proc. Natl. Acad. Sci. USA. 1999; 96:14547–14552. [PubMed: 10588742]
42. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. Blood. 2009; 114:4099–4107. [PubMed: 19706884]
43. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. Philos. Trans. R. Soc. Lond. B Biol. Sci. 1994; 345:101–118. [PubMed: 7972351]

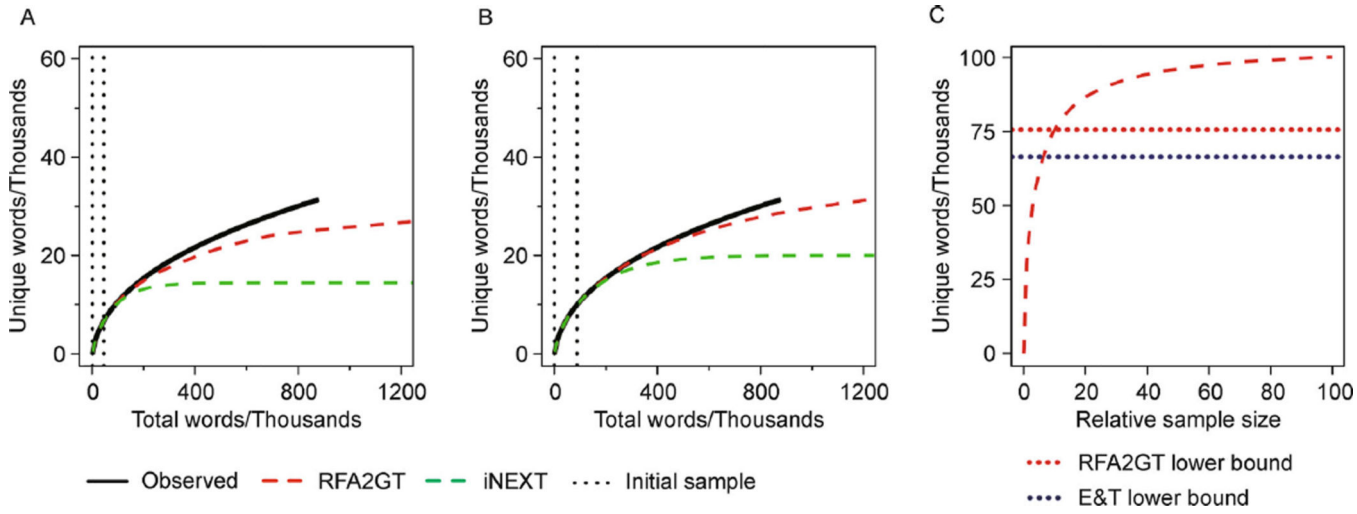

Figure 1. Predicting the number of unique words as a function of the size of the sample

The observed curve is the accumulation curve of the total word counts of Shakespeare's known works compared to the predicted curve (A) when the size of the initial sample is 5% of the total words from Shakespeare's known works; (B) when the size of the initial sample is 10%; and (C) when the size of the initial sample is 100% and comparing the RFA-GT lower bound to the lower bound of Efron and Thisted (E & T).

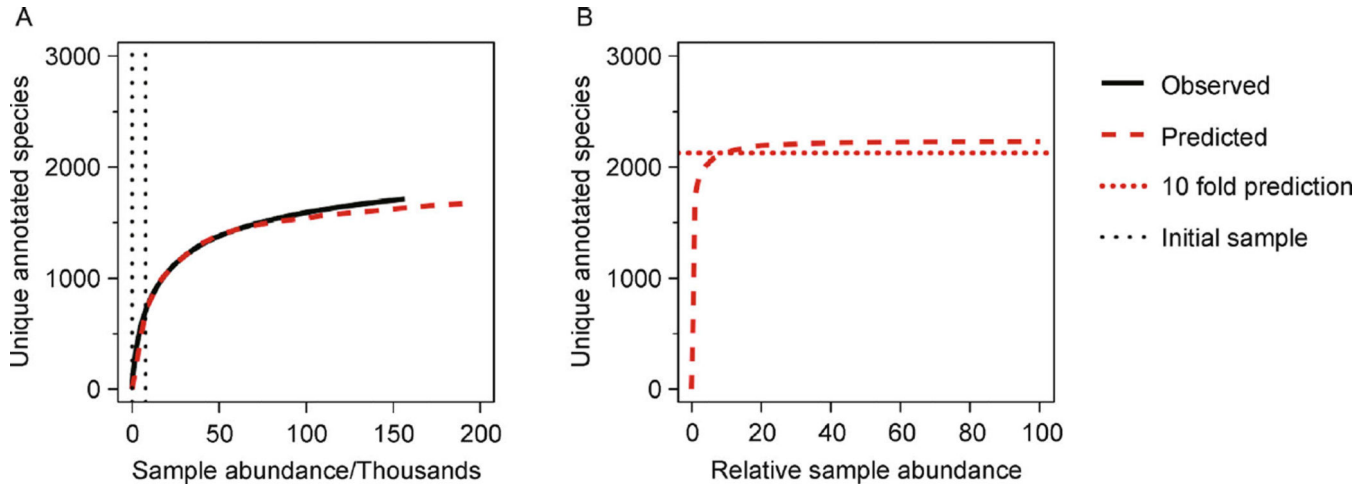

Figure 2. Annotated species as a function of the sample abundance using (A) a 5% subsample and (B) the full observed experiment

The x-axis is the sample abundance and the y-axis is the expected number of unique annotated species. Note that the original curve provided by MG-RAST uses the number of sequences as its x-axis. We convert it to the sample abundance by rescaling.

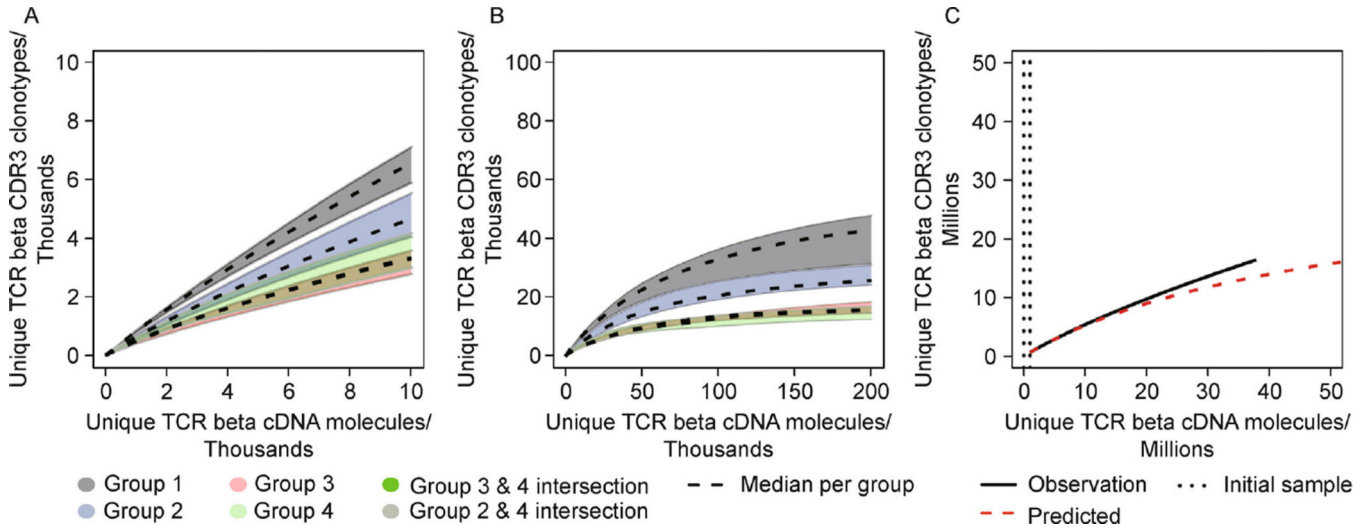

Figure 3. Age-related decrease in TCR repertoire

(A) Interpolation of the accumulation region of TCR β CDR3 clonotypes for each group. (B) Extrapolation of the accumulation region of TCR β CDR3 clonotypes for each group using the observed data in **A**. (C) Predicting the total number of unique TCR β CDR3 clonotypes as a function of the total number of TCR β cDNA molecules by combining all groups.

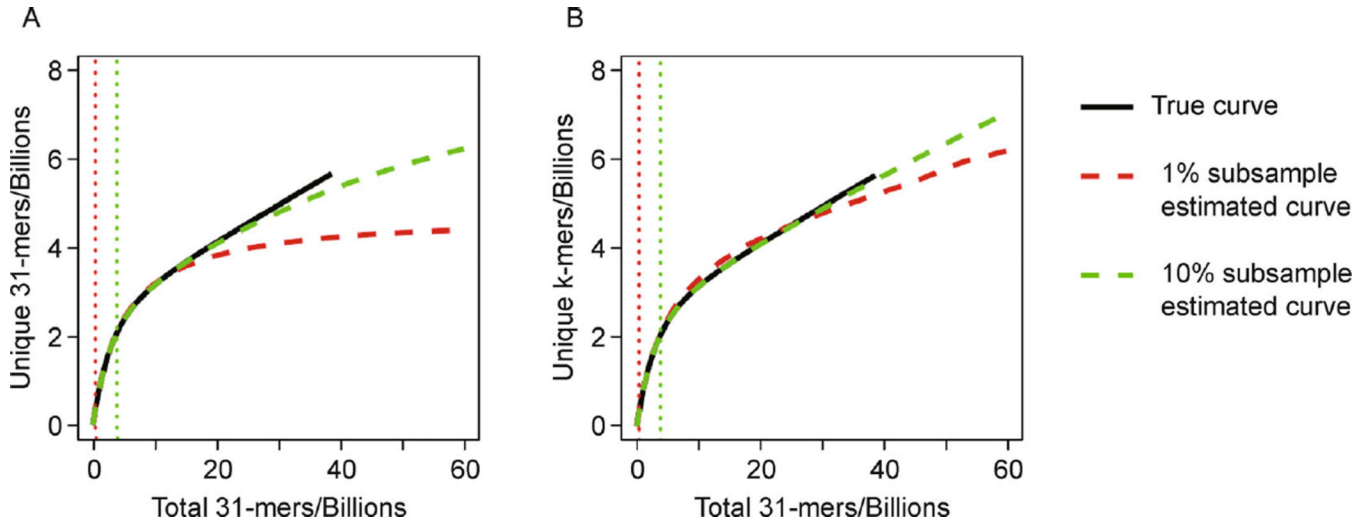

Figure 4. The number of distinct 31-mers as a function of sequenced 31-mers with extrapolations using 1% and 10% subsamples

(A) Extrapolations from the subsamples using default preseqR, with the rational function approximations to the Good-Toulmin power series behaving like a constant asymptotically. (B) Extrapolations from the subsample using rational function approximations to the Good-Toulmin power series behaving like a linear function asymptotically.

Table 1

Frequency of words observed in Shakespeare's known works [4].

j	1	2	3	4	5	6	7	8	9	...
n_j	14376	4343	2292	1463	1043	837	638	519	430	...

Table 2

Predicting the number of unique words as a function of the size of the survey based on 5% sample.

RS	2	4	6	8	10
TV	10017	14470	17759	20452	22771
P, E (%)	9996, −0.2	14064, −2.8	16830, −5.2	18822, −8.0	20614, −9.5
I, E (%)	9830, −1.9	12806, −11.5	13885, −21.8	14277, −30.2	14419, −36.7
RS	12	14	16	18	20
TV	24827	26688	28395	29978	31458
P, E (%)	22035, −11.2	23224, −13.0	24198, −14.8	24786, −17.3	25276, −19.7
I, E (%)	14470, −41.7	14489, −45.7	14496, −48.9	14498, −51.6	14499, −53.9

RS is the relative sample size to the initial survey. TV is the true expected number of unique words observed. P, E is the estimated number of unique words observed by RFA-GT and the corresponding scaled error. A negative scaled error indicates the method underestimated the value. I, E is the estimated number of unique words observed by iNEXT and the corresponding scaled error.

Table 3

Predicting the number of unique words as a function of the size of the survey based on 10% sample.

RS	2	4	6	8	10
TV	14470	20452	24827	28395	31458
P, E (%)	14478, −0.2	20368, −0.6	24191, −3.2	27008, −5.9	29160, −8.7
I, E (%)	14229, −1.7	18111, −11.4	19424, −21.8	19867, −30.0	20017, −36.4

For notations, please refer to Table 2.