ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Molecular Heterogeneity in Large-Scale Biological Data: Techniques and Applications

Chao Deng,[1,*] Timothy Daley,[2,*] Guilherme De Sena Brandine,[1] and Andrew D. Smith[1]

[1]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, California 90089, USA; email: andrewds@usc.edu

[2]Department of Statistics and Department of Bioengineering, Stanford University, Stanford, California 94305, USA

Annu. Rev. Biomed. Data Sci. 2019. 2:39–67

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-072018-021339

*These authors contributed equally to this article

Keywords

library complexity, sequencing coverage, genetic diversity, immune repertoire, single-cell sequencing, saturation

Abstract

High-throughput sequencing technologies have evolved at a stellar pace for almost a decade and have greatly advanced our understanding of genome biology. In these sampling-based technologies, there is an important detail that is often overlooked in the analysis of the data and the design of the experiments, specifically that the sampled observations often do not give a representative picture of the underlying population. This has long been recognized as a problem in statistical ecology and in the broader statistics literature. In this review, we discuss the connections between these fields, methodological advances that parallel both the needs and opportunities of large-scale data analysis, and specific applications in modern biology. In the process we describe unique aspects of applying these approaches to sequencing technologies, including sequencing error, population and individual heterogeneity, and the design of experiments.

1. INTRODUCTION

Library complexity: a measure of the quality of a sequencing library, typically quantified by the number of distinct reads obtained for a given level of sequencing

Genome coverage: the fraction of the base pairs in the reference genome covered by at least one mapped read

High-throughput sequencing is a set of technologies that allow for sampling and identification of the DNA sequence of millions of molecules from a large pool of fragmented DNA. We refer to the pool of fragmented and amplified DNA molecules ready for sequencing as a sequencing library. How these molecules are constructed determines the assay and what the researcher can investigate. There are hundreds of such assays, and an incomplete list is available in Reference 1. Some standard assays include whole-genome sequencing (WGS) and RNA sequencing (RNA-seq). These assays have recently been extended to the single-cell level, which allows researchers to directly investigate cell-to-cell heterogeneity, among other applications.

Although these assays have dramatically increased understanding of many biological principles, there is a hidden hazard inherent to any sampling-based technology. This is the fact that the observed sample is not perfectly representative of the full library, and is often a very poor representation. Rare and low-abundance molecules are hard to sample. Unless the researcher can sequence the sample to complete saturation, there is always a small chance that some molecules are missing from the sequenced reads.

This problem is universally recognized in the statistics literature as the species sampling problem (2), and it dates back to the work of such luminaries as R.A. Fisher (3) and Alan Turing (4). The field has long recognized problems introduced by the missing species when finite samples are the basis for analysis of populations. For example, the maximum likelihood estimate for the frequency of a particular species is proportional to the number of sampled individuals of that species. This implies that any unobserved species will have an estimated abundance of zero. This is a preposterous estimate because without complete sampling there are almost always unseen species in the population, and this estimate can have negative implications for downstream analysis.

In this review, we have two aims: (*a*) to introduce and review statistical approaches that can be applied to analyze molecular heterogeneity and (*b*) to introduce a broad set of modern applications. The applications we cover are intended to illustrate techniques for addressing such problems, highlight their prevalence, and present solutions that may be applied across other applications.

The statistical methodology we cover is rooted in capture–recapture and species sampling statistics and constitutes a rich body of approaches that have yet to be fully explored in the context of modern, large-scale data analysis. In particular, this review covers methods for estimating the species accumulation curve (SAC), species richness, and sample coverage. In Section 2, we define these quantities and provide a review of the theory and methods, although not comprehensively. Classical applications of species sampling involve data sets on the order of hundreds or thousands of captured individuals, and classical methods focus on simple deviations from homogeneity. In contrast, applications in modern biology involve millions or even billions of sampled molecules. This creates new problems but also new opportunities because methods are required that can use the scale of the data to better model the underlying heterogeneity and to make more accurate and powerful inferences. We try to focus on newer methods that are more applicable to large-sampling experiments.

We next explain a set of application areas that share similar questions but have unique features. First, in Sections 3 and 4, we discuss library complexity and the issues of genome coverage in high-throughput sequencing. We focus on estimating how certain summary metrics (e.g., the number of unique molecules and per-base coverage) change as a function of sampling efforts, which relates to SACs. In Section 5, we discuss estimating the genetic diversity across a population: How many genetic variants can be observed in a population as the sequencing effort increases, in terms of both sequencing depth and the number of individuals investigated? We then proceed to discuss the application of species sampling models in the interrogation of the immune repertoire. We examine sampling issues with single-cell sequencing and show how sample coverage could screen

out samples that may not be repeatable. Finally, we briefly discuss difficulties in estimating sample coverage in metagenomic sequencing data. In Section 6, we highlight the effective use of statistical tools from the species sampling problem to high-throughput sequencing applications, and in Section 7, we outline best practices for using these classic methods.

2. STATISTICAL FOUNDATIONS

There are two basic species sampling models that can be considered, the multinomial and the Poisson. In high-throughput sequencing, the sampling depth can be controlled to some extent but is still random, justifying the Poisson assumption. In general, we assume that there are S species with Poisson rates or abundances of $\lambda_1, \ldots, \lambda_S$ and relative abundances of $\pi_1 = \lambda_1 / \sum_i \lambda_i, \ldots, \pi_S = \lambda_S / \sum_i \lambda_i$. Let x_i denote the count of species i and let $N = \sum_i^S x_i$ be the total number of sampled individuals. The underlying assumption is that x_i is a Poisson random variable with $\Pr(X_i = j) = \lambda_i^j / j! \exp(-\lambda_i)$. In all of the cases, we assume that the ordering of the species holds no special significance so that the data can be summarized by the count frequencies $n_j = \sum_{i=1}^S 1(x_i = j)$, the number of species observed exactly j times. For the ease of derivation, it is standard to assume that the rates arise from a common latent distribution $\mu(\lambda)$. Then the expectation of n_j is equal to

$$\mathrm{E}[n_j] = \sum_{i=1}^{S} \lambda_i^j / j! \exp(-\lambda_i) = S \int_0^\infty \lambda^j / j! \exp(-\lambda)\, \mathrm{d}\mu(\lambda).$$

The underlying problem, and the source of all frustration, is that we do not observe the zero counts, n_0. A further complication is that these are necessarily the rarest species, so that the left tail of the abundance distribution is hidden from the investigator, creating what can be considered a length bias in the sampling process because the rarest species are less likely to be samples (5).

We begin by considering a set of related questions that an experimenter might ask about either a sample, a population, or the benefit of additional sampling. In most experimental contexts, the size of the population is unknown and the statistical properties of the population are poorly understood. These questions can be generally summarized into the following categories.

- SACs: If the experimenter samples additional individuals, how many previously unobserved species will be seen? We can rephrase this question in the context of modern sequencing applications as asking how many additional molecules should the experimenter expect to observe from additional sequencing.
- Frequency estimation: What is the expected population frequency of a species observed k times in the sample? To see the importance of this question in data analysis, consider that if the sample does not contain all the species in the population, then the relative proportions of those that have been observed will be on average overestimates of the true proportions.
- Saturation: What is the total population frequency of species represented in the sample and how does this change with additional sequencing? Intuitively, this informs the experimenter about the "weight" of the unobserved species.
- Species richness: How many unique species exist in the population in total? Stated in the language of large-scale biology, this question might be phrased as asking how many distinct molecules exist in a population.
- Notions of diversity: How can a researcher properly compare different populations? In heterogeneous populations, it is not just the number of molecules but also their abundances that distinguish populations.

2.1. Species Accumulation Curve

The SAC is defined as the number of species observed as a function of sampling effort. There are in general two categories of methods to estimate the SAC: to estimate the latent abundance distribution $\mu(\lambda)$ and then calculate the curve under the estimated $\mu(\lambda)$, and to estimate the SAC directly and avoid estimating $\mu(\lambda)$. The first category of methods is split into parametric and nonparametric methods. The parametric methods assume a parametric form for the latent distribution $\mu(\lambda)$. As a notable early example, Fisher (3) assumed that relative species abundance followed a latent gamma distribution and then derived the log-series distribution as the limit as the number of species goes to infinity (6). Many other parametric distributions have been investigated, such as the power-law distribution, the log-normal distribution, and the inverse Gaussian distribution and its generalizations (7–10). Nonparametric methods can be further split into two subclasses, those that attempt to recover the latent distribution via maximum likelihood (11, 12) and Bayesian methods based on the Poisson–Dirichlet process (13). In the former case, since the observed data are discrete, the nonparametric maximum likelihood estimator (NPMLE) is a discrete distribution (14), although one can generalize this to other types of mixtures (15). The Bayesian methods assume that there is an infinite number of species, since if there is only a finite number of species the Poisson–Dirichlet process results in counts that are negative binomial distributed (16).

There are also methods to predict the SAC that do not estimate the latent abundance distribution. For example, Good & Toulmin (17) derived a nonparametric power series estimator for the SAC that performs extremely well when extrapolating to at most twice the original sample size ($t \leq 2$). Unfortunately, the power series diverges rapidly in practice for $t > 2$ due to dependence on the largest observed count (e.g., the highest abundant species). Efron & Thisted (18) constructed estimators based on the Good–Toulmin power series using linear programming and Euler's series approximation. We have proposed a rational function approximation to the Good–Toulmin power series (19, 20). The choice of the rational function allows for both close approximation of the Good–Toulmin power series for small t and stable predictions for large t. It has been hypothesized that the furthest out one can extrapolate the SAC is $t = \mathcal{O}(\log N)$ times the initial sample, $\mathcal{O}(N \log N)$ in total (21). There are two key properties that help in extrapolating the SAC: It must be nondecreasing and convex. Critically, this implies that if two curves cross, then they will never cross again (e.g., see figure 1 in Reference 19), which can help in comparing curves or populations.

Unlike methods that estimate the latent abundance distribution, it is not trivial to extend methods that estimate the SAC directly to predict the number of species represented at least r times (r-SAC) as a function of sampling effort, since both the total number of species and the latent abundance distribution are unknown. One reason is that these curves are not strictly convex and have a single inflection point. Predicting this inflection point is difficult (22). Naturally, one can estimate the underlying distribution and then estimate the curve analytically from the estimated distribution (23). An alternative approach was proposed by Deng et al. (24), who showed that the r-SAC can be derived as a function of the derivatives of the original SAC. Therefore if one can obtain a smooth and accurate estimator of the SAC, then the r-SACs follow directly for any r.

Recently, there has been an interest in the problem of estimating the SAC for multiple samples, and the Good–Toulmin power series can be extended to this case (25). The ideal application of this method would be for cell type identification in single-cell sequencing (26), but in the case of de novo identification of cell types, there are large practical issues with this methodology, as we discuss in Section 5.3.

2.2. Frequency Estimation and Saturation

Frequency estimation is a common application for species sampling models in linguistics. The objective is to smooth the observed frequencies to account for the unobserved species or words. As with the SAC, one can make parametric assumptions about the underlying abundance distribution, use the observed counts to estimate the distribution, and then smooth the frequencies to their expected counts. This includes power law smoothing, of which a special case the well-known Zipf's law (27).

Perhaps the most widely known method of frequency estimation is the empirical Bayes (28) estimator of I.J. Good (29), originally suggested by Turing during World War II (30), and hence known as the Good–Turing estimator. They showed that the expected abundance of a species observed j times is proportional to the fraction of species observed $j + 1$ times. Most importantly, this implies that the abundance of species that are unobserved is proportional to the fraction of singletons in the population:

$$\mathrm{E}\left[\sum_{i=1}^{S} p_i 1(x_i = 0) \,\middle|\, n_1, n_2, \ldots\right] = \frac{n_1}{N}. \quad 1.$$

The abundance of the observed population is commonly called the sample coverage. Equation 1 implies that the sample is completely saturated (all species have been sampled) only if the number of singletons is zero, but this very rarely occurs in any high-throughput experiments.

It may be of interest to extrapolate the sample coverage, similar to the case of extrapolating the SAC (31, 32). This will allow researchers to decide the level of sequencing that achieves effectively no more information (e.g., 33). Indeed, we have shown that the derivative of the SAC is equal to the sample coverage curve (22), so that these two problems are equivalent and all of the methods discussed in the previous section can be applied.

2.3. Species Richness

As in the discussion above, one can always choose between parametric or nonparametric approaches to estimate the number of missing species. In the case of parametric estimation, it is common to apply model selection using AIC (Akaike information criterion) or BIC (Bayesian information criterion) (34), but it has yet to be determined if this form of model selection leads to accurate inferences about the number of species. Even in the case where it is known that all species are equally abundant, it is possible that no unbiased estimator exists (35). In the heterogeneous populations that we encounter in modern sequencing data, parametric assumptions usually grossly underestimate the heterogeneity, which results in an underestimation of the number of missing species. We have found that nonparametric approaches tend to perform better, but these methods also have significant drawbacks, and there is no established method that is superior in all situations.

Estimating the number of species in the nonparametric setting is extremely difficult, as there is an infinite number of possible distributions that can explain a finite sample (36, 37). If the observed counts are assumed to be generated by a Poisson sampling process, then in theory the underlying distribution is identifiable (38), but in practice this is only possible when the sample size is infinite (39). The tactic usually employed to solve this is to limit the space of the underlying distributions.

The most common method is to fit a mixture of discrete distributions and then estimate the total number of species. In general a unique maximum likelihood estimator exists (40). This discrete distribution (NPMLE) (11) will have the number of support points equal to the number of unique species counts. A major issue with the NPMLE is that to guarantee the existence of a

solution, the parameter space must be compact. Therefore an abundance of zero must be included as a possibility. This creates an instability in the estimator and the possibility that an infinite number of species is estimated (41). To solve this issue, researchers have proposed a number of strategies, including penalizing the likelihood to prevent very small abundances (42) and setting a minimum cutoff to define what abundances are interesting (5, 43), since one might consider extremely rare species to be uninteresting in some applications. Another approach offered constructs a discrete distribution by linear programming (44). Alternatively, one can fit mixtures of nondiscrete distributions, such as exponential or gamma distributions (15). However, we have found that these methods have difficulties scaling to sequencing data because of the computational load.

To avoid the problems involved in fitting the underlying distribution of species abundances, researchers have proposed several methods to estimate species richness without estimating the abundance distribution. To avoid explosion of the estimator, one can focus on either strict lower bounds (45) or approximate lower bounds (39). The jackknife estimator is a method developed for correcting these sample size dependencies (46). We have found that this estimator performs well when the sample size is smaller than the number of species but tends to overestimate the species richness when the sample size is larger than the number of species and the abundance distribution is highly heterogeneous (39). Another approach is to estimate the coverage and then estimate the number of species via a Horvitz–Thompson-type estimator (47). In the equal-abundance case, the estimator is known as the Good estimator (29). Chao & Lee (48) added an additional parameter to account for overdispersion, and Chao & Bunge (49, 50) have extended it to the negative binomial model, although this sometimes results in negative estimates of the number of species (39, 42). A new approach is to relate the ratio of successive frequencies to the number of unobserved species, using nonlinear regression to estimate the latter (51, 52). One advantage of this approach is that it can be modified to work without using singletons in cases where the sequencing error is high and the estimated number of singletons in not trustworthy.

Another common approach is to estimate the SAC and then use the predicted number of distinct species for as the sample size becomes large or tends to infinity (20, 53, 54). Unless the estimate is the limit of the SAC as the sampling goes to infinity, this approach tends to underestimate species richness. In general, using extrapolating SACs to estimate species richness is discouraged as unsound statistical practice (55) because the SAC is a function of the sampling effort, which is correlated with, but not predictive of, the true species richness.

2.4. Diversity

Due to the difficulties in estimating the total species richness, researchers have looked to other estimators of diversity that take into account not only the number of species but also their abundances. Some researchers (e.g., 56–58) focus on entropy, usually the Shannon entropy. Another popular diversity index is the Gini–Simpson index (e.g., 59, 60), which is equal to the probability that two randomly sampled molecules are identical.

Due to the influence of unsampled species, the estimates of entropy or diversity will be biased and will depend on the sample size. Many different estimators have been proposed to correct for this problem, including those suggested in References 44, 61, and 62. We have not evaluated these estimators; therefore, we are unable to give guidance on the relative benefits and pitfalls of each estimator.

Both Shannon entropy and the Gini–Simpson index can be generalized as Hill numbers (43),

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}, $$

where $q \geq 0$ and $q \neq 1$. There is another widely used diversity index called Renyi entropy, which is a transformation of Hill numbers (63). It may be desirable to estimate the full range of diversity indexes for all available values of q, but this will be equivalent to estimating the underlying abundance distribution (64). Therefore the same issues that we discussed above in estimating the abundance distribution (e.g., identifiability) also apply here.

As discussed by Kaplinsky and Arnaout (43), no single measure of diversity can capture all features of the populations that one might find interesting. For example, species richness quantifies the number of species and ignores differences in abundances. In the extreme opposite case, the Berger–Parker index measures only the abundance of the most abundant species and ignores differences in the number of species. Diversity measures in between these two extremes quantify to some degree the tradeoffs between the number of species and their relative abundances, and it is up to the researcher to decide which measure is important for their investigation.

3. SEQUENCING LIBRARY COMPLEXITY

The concept of library complexity is best conveyed by providing examples rather than a specific definition. Low-complexity sequencing libraries are those that contain either few unique molecules or a small fraction of unique molecules that appear in many copies in the sequencing library. High-complexity libraries have a large number of unique molecules in approximately uniform abundance. High-complexity libraries are usually preferred by researchers because they maximize the information gained for a fixed amount of sequencing. The practical issues of measuring library complexity also depend on the application, and how to measure unique molecules should be chosen based on the objective of the experiment (see the sidebar titled Counting Unique Molecules).

Evaluating the complexity of sequencing libraries is important for several reasons. First, for problematic samples that may contain a multitude of biases, it is often desirable to avoid the unnecessary cost of sequencing a low-complexity sample or to tailor the level of sequencing based on a cost–benefit analysis, which measures the efficiency of gaining new molecules as sequencing continues. It may also be desirable in such cases to make multiple libraries and choose the most promising one to sequence deeply (**Figure 1*a***). In these cases, a shallow initial sample is sequenced for each sample and then the SAC (called the complexity curve in this case) is estimated to predict the benefit of additional sequencing (e.g., **Figure 1*b***). This allows researchers to spend resources in the planning stage of the experiment in order to save resources (e.g., time, money) when

COUNTING UNIQUE MOLECULES

In any application, applying the approaches covered in this review requires a definition of when two observations represent the same original molecule. In the context of sequencing library complexity, we must determine an identity for each read (or paired-end fragment). This is commonly done by mapping locations in the genome or, more effectively, with unique molecular identifiers (UMIs) (65). Depending on the sequencing application, if two reads map to the same location in the genome, we might have evidence that they originated from the same molecule [e.g., from the same PCR (polymerase chain reaction) clone]. In applications where reads are expected to map to very precise positions, for example, in targeted sequencing (66), enrichment-based protocols (67), or droplet-based single-cell RNA-seq protocols that capture only one end of the transcript, the expected coverage for the desired regions will be extremely high, and even reads from distinct molecules often map to the same location. UMIs provide additional information in the form of random barcodes ligated to molecules prior to PCR amplification.

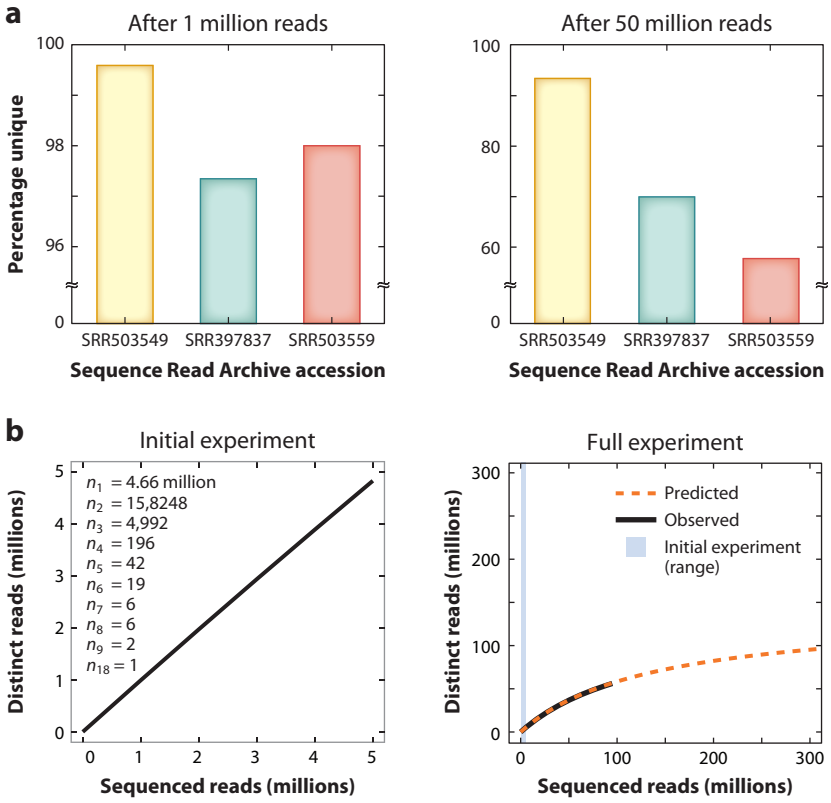

Figure 1

(*a*) The number of distinct reads for (*left*) an initial sample of one million (M) reads and (*right*) a full sample of 50 M reads in three human ChIP-seq (chromatin immunoprecipitation and sequencing) experiments (70, 71). (*b*) To determine the desired sequencing depth for a bisulfite sequencing sample of gorilla sperm, we performed an initial shallow sequencing sample of approximately 5 M reads (*left*) and used this to predict the complexity curve (*right*) to decide the desired sequencing depth. We chose a depth of approximately 100 M reads and obtained 56.1 M distinct reads compared to the prediction of 56.6 M (72).

conducting the full-scale stage of the experiment. It also allows researchers to compare and evaluate technologies in a comprehensive manner (66, 68, 69). Other researchers have used complexity curves to compute the desired sequencing depth to achieve near-saturation of the sample (20, 33), but, as we discussed in Section 2.2, there are principled ways to estimate the saturation to guide such decisions.

4. GENOME COVERAGE

Genome coverage is one of the key elements in the design of high-throughput sequencing experiments (73). High coverage can avoid sampling bias and help to make robust biological discoveries. Researchers investigating samples that have small amounts of DNA, as a result of either degradation or low number of cells, require high amounts of amplification, and coverage is often a major issue. One example is single-cell DNA sequencing, where coverage varies wildly across amplification methods (74). It is possible that some regions of the genome are missing entirely from the sample and will never be covered, no matter the depth of sequencing (**Figure 2**).

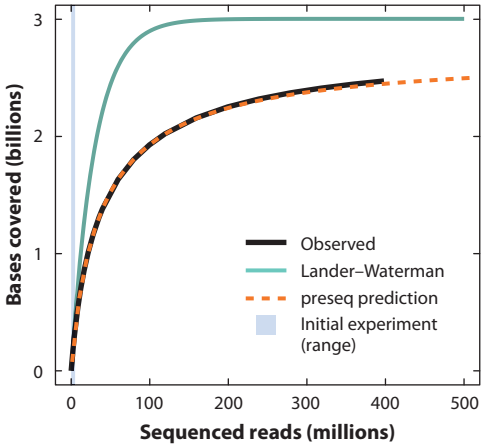

Figure 2

Observed coverage of whole-genome sequencing of a single cell (75), along with the expected coverage under the theoretical Lander–Waterman model (76) and preseq predictions (74) from an initial sample of five million reads.

There are several ways to quantify sequencing coverage (see the sidebar titled Counting Strategy). The naïve measurement is the number of reads, commonly used in RNA-seq and other sequencing applications where the reference is either not available or incomplete. When the reference is available, e.g., in human WGS, coverage is conveniently measured by the average sequencing depth, $\lambda = NL/G$, where N, L, and G are the number of aligned reads, the length of a read, and the size of the reference, respectively. This metric reflects the average number of reads that cover a random position in the reference genome. For example, a sample with 30× mean coverage for human WGS means that each base pair in the genome is on average covered by 30 sequenced reads. Note that the average sequencing depth contains little information about the evenness of coverage along the genome. Since reads are not evenly distributed along the genome, it is possible, and commonly observed, that the majority of reads are aligned to only a small fraction of the entire genome, leaving the other regions uncovered. In this scenario, the average sequencing depth does not reflect the actual coverage of the genome. An accurate measurement is per-base coverage, which describes, for each nucleotide in the reference genome, the number of reads that

COUNTING STRATEGY

Classic estimators require summary statistics to predict the SAC (77). The summary statistics are represented by a two-column table. The first column is the frequency, j, and the second column is the number of species sampled j times in the sample, known as the frequency of the frequency j. The meaning of SACs depends on the definition of the species and the counting strategy. For example, to predict library complexity, one should consider a unique read as a species, and the frequencies of the species are counted by the duplicates of the read (19). The estimated SAC should be interpreted as the number of unique reads versus sequencing efforts. When the purpose of sequencing is related to genome coverage, as in applications focused on genotype, it is appropriate to use base pairs in the genome as species, and the frequency of a species is counted by the number of mapped reads that cover the base pair (74). In this case, the estimated SAC should be interpreted as the number of base pairs covered in the reference genome versus sequencing efforts.

cover it. For example, the Broad Institute Genomic Service sets a gold standard for human whole-exome sequencing (WES) at 85% of the exome covered by at least 20 reads (**https://genomics.broadinstitute.org/products/whole-exome-sequencing**).

Lander–Waterman curve: the expected coverage under the assumption of equal probable sampling of all possible molecules

Coverage requirements depend on the specific application, the size of the regions of interest, the read length, and so on. For example, Pollen et al. (78) showed that shallow single-cell RNA-seq (scRNA-seq) (≈50,000 reads per cell) is sufficient to categorize cell types, and recent advances in single-cell technology have moved toward sequencing hundreds of thousands of cells at very low coverage (79). Genome assembly typically requires extremely deep sequencing to accurately reconstruct the whole genome (80). ChIP-seq (chromatin immunoprecipitation and sequencing) coverage is determined by the particular epigenomic mark or transcription factor of interest (81); for example, higher coverage is required for histone modifications organized as broad domains, compared to narrow peaks for other histone modifications or transcription factors. Most users determine the necessary coverage based on personal experience, empirical evidence collected from literature, commercial recommendations, and best practices defined by the scientific community (82). For example, Sims et al. (73) provided general guidelines for sequencing coverage for various applications. While these guidelines are very helpful, there are issues with this strategy. General guidelines tend to be conservative and following them may waste resources. Many factors from both experimental artifacts and the stochasticity inherent to each biological system could affect coverage of sequencing data (73). It is impossible to find general guidelines that can be appropriately applied in all cases. One valuable question is, When sequencing data appear insufficient, how many additional reads should be sequenced?

The pioneering work in this area is the well-known Lander–Waterman theory (76), which estimates per-base coverage in a sample for Sanger sequencing (83). The Lander–Waterman theory has been widely used by various sequencing projects and is suggested by Illumina (82) as a guideline to estimate the desired sequencing coverage. The basic statistical assumption is that reads are generated uniformly at random from the genome. Under this assumption, if G is the known size of the genome and $\lambda = NL/G$ is the average sequencing depth, then the expected number of base pairs covered in a sample is estimated as

$$G[1 - \exp(-\lambda)], \qquad 2.$$

known as the Lander–Waterman curve (see the sidebar titled A Next-Generation View of the Lander–Waterman Model). For single-nucleotide polymorphism (SNP) detection, it may be more interesting to estimate the expected number of base pairs that are covered by at least r reads, which is equal to

$$G\sum_{i\geq r}\lambda^i/i!\exp(-\lambda) = G\left[1 - \sum_{i=0}^{r-1}\lambda^i/i!\exp(-\lambda)\right].$$

Here, $r > 1$ is the minimum required coverage to call a SNP.

The assumption that reads are uniformly represented in the library and across the genome, known as the homogeneity assumption, may not account for the heterogeneity of the abundances of reads in high-throughput sequencing. For example, GC- and AT-rich DNA fragments are underrepresented in sequencing results (85). To account for heterogeneity, one typically assumes that reads arise from a Poisson mixture. Similar to the homogeneous model, the mixture model assumes that for a given sample, coverage of each base pair follows a Poisson distribution. Instead of using the same rate λ for every base, the mixture model assumes that the rate λ_i for each base pair i is different and is generated from another latent distribution $\mu(\lambda)$. In particular, if the

A NEXT-GENERATION VIEW OF THE LANDER–WATERMAN MODEL

In the Lander–Waterman model, clones are grouped as islands comprising one or more clones, based on their fingerprints (we are ignoring the concept of so-called apparent islands, which concerns the overlap criteria between clones in an island). Islands with at least two members are called contigs (84). Among the most important questions answered in Lander and Waterman's seminal paper (76) are the expected number of islands comprising k clones, the expected number of clones in an island, and the expected size (in base pairs) of an island. Their methodology worked in the context of sequencing clones because the assumption of a Poisson distribution was sufficiently accurate. We might consider adapting the methods outlined in this review to answer analogous questions for contemporary sequencing technologies. This would require some means of identifying sequenced reads (instead of clones), with other reads grouping them as islands. If we imagine sequencing a small number of reads, we can test pairs of reads to determine if they meet the criteria for belonging to the same island. This will eventually fail because all the methodologies discussed in this review assume that the number of species (in this application, islands) never decreases. However, after sufficient sequencing, we might learn that the islands begin to merge, and our SAC would adopt a negative slope.

mass of $\mu(\lambda)$ concentrates on a single point, the mixture of Poisson distributions degenerates to a homogeneous model. As an example, consider RNA-seq. The difference of abundances between highly expressed genes and lowly expressed genes could be orders of magnitude. Transcripts from highly expressed genes are therefore more abundant in the sequencing library and more likely to be sequenced than transcripts from lowly expressed genes. A common assumption is that the latent distribution belongs to the gamma distribution family, $\Gamma(\alpha, \beta = \alpha/\lambda)$, where α and β are parameters to determine the gamma distribution and λ represents the average sequencing depth (86–88). Then the number of reads covering a base pair follows a Poisson–gamma distribution, also known as a negative binomial distribution. In this case, the expected number of base pairs covered in the sample is estimated as

$$G\left[1 - \left(\frac{\alpha}{\alpha + \lambda}\right)^{\alpha}\right].$$

As an alternative, Daley & Smith (74) proposed a nonparametric method to solve the problem without inferring the latent distribution. This method shows promising results on whole-genome single-cell sequencing experiments where read abundances are highly heterogeneous due to the amount of amplification required. However, because this method does not infer the latent distribution, it is not trivial to predict the number of base pairs that are covered by at least r reads in the sample. Deng et al. (24) discovered a relation between the number of base pairs covered by at least r reads and the number of base pairs covered under the mixture of Poisson distribution. They showed that the number of base pairs covered by at least r reads in the sample can be estimated as a simple rational function expression.

5. ADDITIONAL APPLICATIONS

5.1. Genetic Diversity

Several large-scale sequencing projects have been launched to find human genetic variations in multiple populations. For example, The 1,000 Genomes Project, which began in 2008, has analyzed 2,504 genomes across 26 population (89). The project found over 88 million variants,

including 84.7 million SNPs, providing a basis for studying genetic diseases (89). Very recently, such strategies have been applied to single-cell data at the individual level to determine somatic mutation rates and their consequences (90). The aim of these projects is to identify as many genetic variants as possible. However, due to rare genetic variants, limitations of sequencing technologies, and the evolution of mutations, it is impossible to identify all variants. To evaluate the completeness of identified genetic variants and to help design future experiments, it is important to estimate the properties of unseen variants. In this situation, each person or cell is an observation, and then variants are measured for each observation. In this sense, there are two levels of sampling: the number of individuals to sample and how deep to sequence each individual. If we treat as fixed the sequencing depth and the number of possible variant sites (e.g., the length of the genome), then the data can be presented as a list of binary indicator variables, one entry for each variant and one list for each individual. This structure is known in the capture–recapture literature as incidence data (91–93).

Before diving into the problem of estimating the number of variants, we first examine how to detect an allele. Two important components are the number of samples and sequencing coverage. Intuitively, detecting an allele requires multiple samples that contain the allele, with each sample containing many reads that cover the allele. We use a simple statistical model to demonstrate how these two components affect the power of detection. Let θ be the frequency of a randomly selected variant in the population. Based on the Lander–Waterman model, the number of reads that cover a position in the genome follows a Poisson distribution. The Poisson rate λ is estimated by the average depth of sequencing. We use N to denote the number of aligned reads for this individual. To simplify the problem, we ignore sequencing error and define the power of detecting a variant as the probability of having at least one read containing the variant. The probability of having no reads containing the variant is calculated as

$$1 - \sum_{i=1}^{N} \binom{N}{i} (1-\theta)^{N-i} \theta^i \exp(-i\lambda) = 1 - [1 - \theta + \theta \exp(-\lambda)]^N, \quad 3.$$

which is the minimum power of detecting a variant. The power increases as either sequencing coverage or sample numbers increase. When the variant is rare in the population (θ is small), the power of detecting the variant is low. In particular, from Equation 3, the power is less than $1 - (1-\theta)^N$. Therefore, to detect rare variants, it is important to have many samples to increase the power for detection. There are other approaches to improve the power of variant detection. For example, The 1,000 Genomes Project mixed WGS and WES to cover common variants and rare variants from the protein-coding regions (89). In this section, we assume all factors are fixed except for the number of individuals.

Ionita-Laza et al. (92) assumed that the frequency of an allele θ follows a beta distribution. The parameters of the distribution are estimated by maximizing the zero-truncated likelihood of the observed variant frequencies. Instead of using the parametric distribution, Zou et al. used a nonparametric method to estimate the distribution of θ (23). They assumed that θ follows a discrete distribution and constructed the simplest distribution that can fit the data n_j using linear programming (23, 44). Once the latent distribution $\mu(\theta)$ is known, one can calculate the probability P_0 of missing a variant and estimate the total number of variants as $\hat{S} = S_1/(1 - P_0)$. Alternatively, Gravel (93) presented a nonparametric linear programming algorithm for estimating the number of variants as a function of individuals, using ideas presented by Efron & Thisted (94) in the context of linguistics. Gravel et al. (95) proposed a jackknife-based approach to estimate S. The basic idea is to use the number of observed SNPs to estimate S. It is clear that the number of observed SNPs is a lower bound for S given finite number of samples. However, if the bias is of the

form $E[S] = S_1 + c_1/N + c_2/N^2 + \ldots$, where c_i are constants, the jackknife estimator can further reduce the bias up to any given order (96).

One should be cautious when directly applying estimators for species richness to predict the total number of variants, for a number of reasons. First, the problem of estimating species richness is very challenging, as we discussed in Section 2.3. In general, there is no unbiased estimator available for S from the sampling data. Intuitively, no matter how large a sample is, there may still be many variants with low abundances in the population. Even though we can bound the number of possible variants, distinct allele frequency spectra can lead to identical observations (36, 93). Therefore, the estimation of species richness heavily depends on the statistical assumptions concerning frequency spectra in the model (see Section 2.3). Second, calling a variant is more challenging than identifying which species is observed in other applications. Sequencing errors, PCR (polymerase chain reaction) biases, and the quality of the reference genome can cause false positives. As a result, the estimated genetic diversity can easily be an overestimate. Typically, variant calling contains three basic components (97). First, preprocessing filters out reads with low quality. Different methods are then applied to call variants. Finally, postprocessing is further applied to reduce false positives by adding other filters. The basic idea of these methods is to simulate the process of sequencing and to correct artifacts to reduce false positives and negatives. Compared with the theoretical difficulties in estimating species richness, the accuracy of variant calling continues to increase because improvements in sequencing technology keep reducing the impact of many of the above-mentioned artifacts.

5.2. Immune Repertoire

T cells and B cells express receptor proteins (called antigens in the case of B cells) that function in vertebrates as a form of adaptive immunity (98). T cells and B cells both undergo DNA rearrangement during the development of a specific region in the genome consisting of variable (V), diversity (D), and joining (J) gene segments. This process is aptly named V(D)J recombination. Additional processes such as mutations, insertions, and recombination can generate additional diversity. The result is a highly diverse repertoire of immune receptors, with lower bounds of 10^{15} and 10^{18} (99) and upper bounds of 10^{61} and 10^{68} (100) for the total possible receptor diversity for T cells and B cells (100), respectively. Even the lower bounds are much larger than the number of cells in the human body ($\approx 10^{13}$); therefore, not all possible receptors can be present.

Not all possible receptors can be generated and it is unlikely that all are equally likely to be generated. Even if they could be uniformly likely, the body has mechanisms to select for the more beneficial receptors and against others. For T cells, this occurs in the thymus, where immature T cells are tested for their binding affinity to peptides. If the binding is too high or too low, then apoptosis is induced in the T cell, and only the T cells whose binding affinity is just right survive selection and are allowed to exit the thymus (98). Subgroups of cells with a particular V(D)J arrangement are known as a clones. Peripheral T cell division can lead to additional expansion of a particular clone outside the thymus (101). This implies that the distribution of T cell receptor repertoire will be highly diverse and difficult to estimate. Indeed, Mora et al. (102) proposed that a natural model for immune repertoire diversity is a power law due to selective pressure for receptor fitness.

During an organism's lifetime, there are many additional selective pressures that will induce diversity into the immune repertoire. The full mechanisms for this are not fully understood, but it has been shown that immune repertoire decreases with age (**Figure 3**) (50, 103), is negatively correlated with health (104), responds to illness and vaccination (105, 106), and is positively correlated with tumor mutation load in cancer (107). Associating immune repertoire with such external factors can elucidate the underlying relationship between the immune system and the environment.

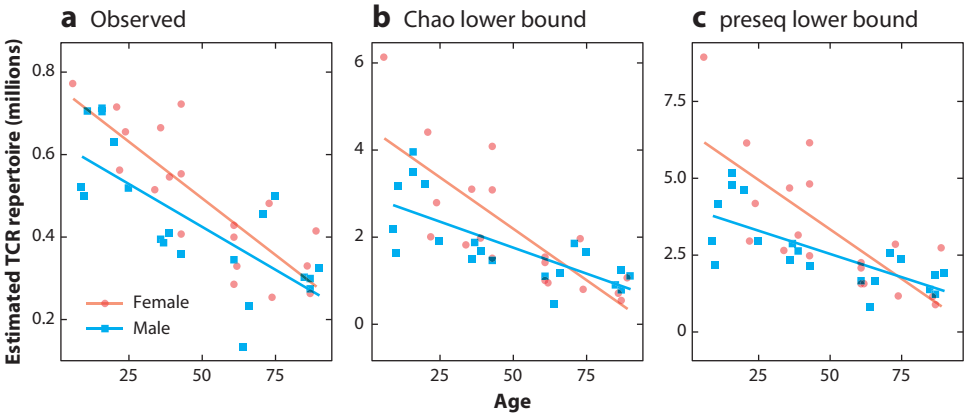

Figure 3

T cell receptor (TCR) total repertoire estimates from (*a*) the observed samples, (*b*) Chao's lower bound estimates (45), and (*c*) preseq lower bound estimates (39), along with linear fits as a function of age and sex for 38 individuals from Reference 103.

The question then becomes how to measure diversity of the immune repertoire. As discussed in Section 2, there are many ways to mathematically define diversity. The naïve way to measure diversity is with the number of observed receptors. As we discussed throughout this review, the total number of unique observed species is highly dependent on sampling or sequencing depth. Furthermore, it ignores any useful distinction between rare versus abundant receptors. The next most obvious way is to estimate the total number of T cell or B cell receptors by estimating the number of unobserved receptors and adding that to the number of observed receptors. This approach can be problematic because the total number of such species is usually extremely hard to quantify. For example, if the underlying population follows a power law, then estimating the total population can be difficult for two reasons. First, the abundance distribution depends strongly on the normalization constant, which is dominated by a very large number of extremely rare species (100), which are not well captured by incomplete samples. Second, even estimating the exponent accurately is difficult in the presence of censored data (108, 109), and the usual method of validating a power law by use of log rank–log frequency plots is problematic (110).

Because of the difficulties in estimating the number of missing species, other measures of diversity have been proposed for estimating and comparing immune repertoire. Some researchers (e.g., 56, 57) focus on entropy, usually the Shannon entropy. Another popular diversity index is the Gini–Simpson index (e.g., 59, 60), which is equal to the probability that two random T or B cells arose from the same clone.

Even exhaustive sequencing at an exorbitant cost is not sufficient to accurately quantify the immune repertoire, as some receptors may not be present in a blood sample but are present in the body (111). This brings up a difficult point. Researchers in general would like to understand the immune repertoire at a individual level, but species models can only make inferences at the sample level. For a single sample, extrapolation or inference of diversity indices can only be extended to the sample population—the repertoire (or molecules) that is contained in the sample and could be observed if sequencing were exhaustive. Extensions to the individual level require multiple samples per individual, resulting in incidence data (91, 112). Modeling this data is more complex, as it requires modeling both the missing species within each sample and the missing species between the samples.

Similar to how the scale of high-throughput sequencing data induced dramatic changes in sample-based inference on species diversity, we believe that the ever-decreasing cost of sequencing-based technologies will enable new, improved models for incidence data.

5.3. Single-Cell RNA Sequencing

Single-cell sequencing has become a popular method to profile the heterogeneity of biological systems that may be hidden by bulk analysis. In particular, recent technological improvements in scRNA-seq have increased the throughput from a few cells to hundreds of thousands. This has allowed researchers to conduct unbiased surveys of complex populations and discover novel cell types/states through the use of computational clustering of expression patterns. It is a fundamental requirement that the per-cell coverage in these data sets be large enough for statistically robust results in downstream analysis, and knowing how much to sequence in a sample becomes a crucial decision to make.

As the cell throughput increases, often scientists must find an optimal trade-off between the number of cells to sequence and the average coverage each cell will have. In the first single-cell technologies (113), the number of cells was small enough that such questions were moot. As barcoding methods started to allow cells to be tagged and multiplexed, data sets saw a large decrease in the per-cell coverage distribution. In such cases, the library comprises not just one cell, but a large and unknown number of cells (see the sidebar titled How Many Cell Types Exist in the Population?). Part of the processing step is to determine the number of cells based on the number of expressed genes and unique molecular identifiers (UMIs) in each cell (see the sidebar titled Counting Unique Molecules).

Droplet-based technologies use barcoding methods to deconvolute the cellular origin of each read (see the sidebar titled Massively Parallel Single-Cell RNA Sequencing). Single cells are captured by droplets that contain a lysis buffer to break the cell and reverse transcriptases that contain

HOW MANY CELL TYPES EXIST IN THE POPULATION?

For many researchers eager to apply massively parallel scRNA-seq, the most appealing possibility is discovering new cell types. These are typically expected to be rare, as otherwise they would already have been discovered. On the surface, it might seem obvious that capture–recapture statistics can be applied to estimate the number of new cell types that would be observed if we sample individual cells more deeply—essentially sequencing more deeply from prepared libraries, or preparing additional libraries from the same tissue samples. Examining this task in more detail reveals specific challenges. To obtain a counts histogram for applying species sampling models, we must associate individual cells with cell types. This is typically done by clustering the gene expression profiles for each individual cell—the clusters correspond to cell types. We may assume an adequate means of identifying clusters, determining the appropriate number of clusters, and that the clustering process is successful. We still have the problem of assigning cells to clusters. Drawing hard cutoffs based on, e.g., distance to cluster centers may lead to ambiguities or artifacts. Using a soft clustering method, like a Gaussian mixture model, we could apply probabilistic clustering, use randomization, and obtain a counts histogram (or repeat the process to obtain many). But another problem emerges at this point. After sampling additional cells, the identities of the original clusters might change (114, 115). Some of the original clusters might merge together, similar clusters may be better distinguished with more cells, and we may obtain fewer or more cell types after sampling a greater number of cells. Indeed, what exactly defines a cell type by its expression signature is still an unresolved question in the field (116). Therefore, at present, species sampling models are not directly applicable to the problem of estimating the number of cell types in a population.

MASSIVELY PARALLEL SINGLE-CELL RNA SEQUENCING

scRNA-seq methods initially required obtaining cells manually (113), which limited their throughput. Microfluidics-based technologies increased the throughput to tens or hundreds of cells (117). Massively parallel single-cell sequencing emerged in 2015 with immediate impact when two papers (118, 119) simultaneously introduced related approaches that had the capacity to conduct scRNA-seq for thousands of cells in parallel using droplet-based technologies. One limitation of droplet-based technologies is that it only captures the 3′ end of the transcript. Automated machines such as 10X Chromium (120) and ddSEQ™ (121) have democratized the application of scRNA-seq, allowing researchers everywhere to investigate heterogeneous populations.

both a cell barcode (CB) and a set of random UMIs that attach to each captured mRNA transcript. All cells are sequenced together through paired-end reads, in which one end will contain the CB and the UMI, while the the other end will contain the 3′ end of the captured transcript. In applications such as RNA-seq, whose aim is to quantify gene expression, it is important to separate reads originating from the same gene that contain different UMIs (thus implying large quantities of that gene or a large gene length) and reads with repeated UMIs, which indicate that multiple PCR-amplified copies of the same read were sequenced. The latter, however, is an important metric of saturation, since the combination CB+UMI is the species observed in the sequencing library.

CBs and UMIs can be repeated, and different RNA molecules may contain both. The currently adopted standard is to consider reads as a product of overamplification if both contain the same CB+UMI and map to the same gene in the genome. In this case, we can summarize the frequency with which each read was observed in the same way as we do with species, and a direct application of species models allows us to interrogate how many new molecules are expected if we increase the amount of sequencing.

A measure of the completeness of the sampling of the transcriptome for each captured cell is the sample coverage (see Section 2.2), a measure of the percentage of the transcriptome sampled, equal to one minus the number of singleton UMIs divided by the total number of identified UMIs (122). This measure will identify insufficiently sampled cells that could be problematic in downstream analysis. Alternatively, one could use species sampling models to estimate the number of undetected genes. These are genes that are in the library and would be observed with deeper sequencing, but not genes that are missing from library. The latter are known as dropout genes and tend to be rare and lowly expressed genes (123). Early methods used a constant rate of dropout across all cells (124), but recently it has been recognized that cell-level estimates of the dropout rate are important for accurate expression quantification (123, 125).

5.4. Future Issues

In the future, species sampling models may be able to illuminate certain issues in scRNA-seq, including:

1. How should researchers distinguish undetected genes from dropout genes based on scRNA-seq data?
2. How should researchers design experiments to optimize discovery for cell type identification, and how can species statistical models help to inform this problem?
3. Rather than simple cutoffs for UMI and CB counts for determining real cells, can species sampling models help determine optimal methods for determining real cells?

OPERATIONAL TAXONOMIC UNIT

Operational taxonomic units (OTUs) are clusters of similar sequences that, in theory, reflect the evolutionary relationship of species (131). OTUs are commonly used in 16S sequencing as a working definition of species. There are two basic strategies to construct OTUs. For reference-based approaches, reads are mapped to cluster centroids, which are chosen from predefined references. Therefore, reads that have low similarity to centroids are excluded from analysis. This strategy does not account for unknown bacterial species. For de novo approaches, reads are aligned against each other and grouped into OTUs. This strategy does not require references but is sensitive to predefined thresholds and the clustering algorithm, and it ignores information on a finer scale such as strain-level differences (132, 133). Although it is popular in 16S sequencing, OTUs are rarely used in whole-metagenome shotgun sequencing due to sequencing errors. Methods that are designed for 16S sequencing data often generate many spurious OTUs, which inflates the diversity of the sampled microbial community (134).

5.5. Metagenomics

Metagenomics is the study of environmental samples through directly sequencing their genetic materials (126). While the traditional way of studying microorganisms has been limited by culturing bacteria in the lab, metagenomics can sequence microbial species in the sampled microbial community regardless the culturability of the species. Several large metagenomic sequencing projects have been carried out, such as the Human Microbiome Project (127, 128) and the Tara Oceans Expedition (129). These projects greatly improve our understanding of the microbiome in our bodies and throughout the planet.

There are in general two ways to investigate microbial species in metagenomics: targeted sequencing, such as 16S RNA-seq, and nontargeted sequencing, known as whole-metagenome shotgun sequencing. The application of species models to the former is thoroughly discussed in Reference 130 and requires methods of distinguishing species, for example, by operational taxonomic units (OTUs) (see the sidebar titled Operational Taxonomic Unit). In this section, we focus on one of the fundamental but difficult problems in whole-metagenome shotgun sequencing: how saturated a metagenomic data set is.

The statistic $1 - n_1/N$ (see Section 2.2), which is an asymptotically unbiased estimator for sample coverage, is probably the simplest metric to measure the saturation. However, in the context of metagenomics, it is not trivial to even obtain an appropriate value of n_1 due to sequencing errors and incomplete reference genomes. For example, we can measure the abundance of a bacterial species only if the bacterial species has a reference genome. In this case, reads that cannot map to reference genomes are filtered out. Therefore, the summary statistics n_j represents the number of annotated bacterial species with frequency j. These summary statistics do not contain information about unknown bacterial species in the sample.

One type of method is to directly use reads to measure the saturation of metagenomic data. Similar to the example above, the difficulty lies in how exactly one should generate the summary statistics n_j for $j \geq 1$. Due to sequencing errors, the raw data contain many false reads that are not from the sequencing library. The naïve way of counting abundances of reads could severely inflate the value of n_1 and largely underestimate the sample coverage. For scRNA-seq, the identity of each read is determined by its UMI (Section 5.3), which is much shorter than a read and therefore is more robust to sequencing errors. If the reference genome is available, one can map reads to the reference genome and then use the mapping location to define a unique read (see Section 3). Because identical reads with a few errors can be mapped to the same location, this approach is robust to sequencing errors. Unfortunately, sample microbial communities typically contain many

unknown species, whose reference genomes are not available. To overcome the problem, one can align reads against one another and set some thresholds to define duplicate reads. This strategy is computationally expensive and sensitive to the choice of thresholds. Rodriguez & Konstantinidis (135) proposed a sampling method to reduce the computation, but the method may be sensitive to sequencing depth.

We have found success with using *k*-mers to measure the saturation for a metagenomic data set. The idea is inspired by de Bruijn graph, which is used to correct sequencing errors for genome assembly (136). A *k*-mer is a substring of length *k* from a read, so that a read of length *L* can be decomposed into $L - k + 1$ different *k*-mers. The basic assumption is that false *k*-mers, those containing sequencing errors, tend to have low frequencies in the sample compared with true *k*-mers. Therefore, the number of abundant *k*-mers, which are considered true *k*-mers, versus sequencing effort should reflect the genuine saturation of the metagenomic data. We use a threshold *r* to distinguish abundant *k*-mers from low-frequency *k*-mers. The larger the value of *r* is, the more robust the curve is to the sequencing error. As a trade-off, a large *r* value requires a large sequencing effort to reveal true *k*-mers.

Figure 4 shows the number of unique *k*-mers represented at least two times as a function of sequencing effort for metagenomic data from various environments. Each sample is extrapolated to 100 billion *k*-mers using preseqR (24). The trend of each curve shows how saturated the metagenomic data is. The orders of the curves indicate the complexity of micro communities from different environments. Note that this strategy may not be applied in a read level. Because the probability of generating a false read should be higher than the probability of a false *k*-mer due to the length, the metagenomic data may not real enough high-frequency reads for constructing estimators.

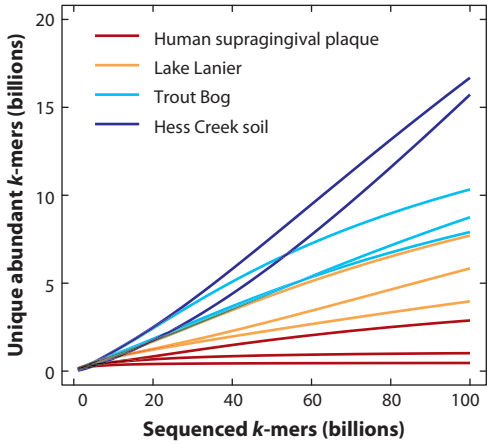

Figure 4

Unique *k*-mers represented at least two times versus total sequenced *k*-mers. We collected 11 public metagenomic data sets from human supragingival plaque (*red*), Lake Lanier (*orange*), Trout Bog (*light blue*), and Hess Creek soil (*blue*). We used jellyfish (137) to obtain the histogram of 31-mers for each data set and preseqR (24) to predict the number of 31-mers represented at least twice as the number of sequenced 31-mers increases to 100 billion. Given the same amount of 31-mers, samples from soil were the least saturated among all samples. In the contrast, human samples were the most saturated. The saturation of the Trout Bog samples was between that of Hess Creek soil and Lake Lanier. The order of saturation is consistent with the order of complexity of microbial environments, which from the most complex to the least complex are soil, bog, lake, and human.

Table 1 Software for species estimation

Package	Description	URL	Reference(s)
SPECIES	Multiple species richness estimators, including Chao's lower bound, ACE, jackknife, and NPMLE	SPECIES R package	138
preseq	Species richness lower bounds, RFA to Good–Toulmin estimator for SAC	preseq C++ package	19, 39
preseqR	r-SAC estimation	preseqR R package	24
CatchAll	Parametric species richness estimators	CatchAll C# package	34
SpadeR	SAC and species richness estimation	SpadeR R package	139
breakaway	Ratio regression estimation of species richness	breakaway R package	51
edgeR goodTuring	Good–Turing frequency estimation based on Reference 140	goodTuring function in edgeR	141
EstimateS	Species richness and shared species estimation	EstimateS software	142
Nonpareil	Sample coverage estimation for metagenomic data	Nonpareil C++ package	135
Recon	Discrete density estimation for immune repertoire analysis	Recon Python package	43

Abbreviations: NPMLE, nonparametric maximum likelihood estimator; RFA, rational function approximation; SAC, species accumulation curve.

6. AVAILABLE SOFTWARE TOOLS AND BEST PRACTICES

Many software have implemented these methods. **Table 1** contains a complete list. In the following sections, we provide some guidance in applying the principles illustrated in the examples above. We first focus on a set of pitfalls to avoid.

6.1. Is the Total Number of Base Pairs Known?

Researchers sometimes use the length of the reference genome, G, to approximate the total number S of base pairs when estimating genome coverage versus sampling efforts (see Section 4). This simplifies the estimation for methods that rely on S but makes no difference for methods that do not infer the latent distribution (e.g., 19).

We use Lander–Waterman theory (Equation 2) as an example to demonstrate the difference of treating S as known versus unknown. If the value S is known, for example, the length of the reference genome G, the parameter λ is estimated as

$$\lambda = \sum_{i \geq 1} i n_i / S$$

based on maximum likelihood. If the value of S is unknown, so does the zero counts, n_0. The parameter λ then is estimated by maximizing the likelihood of the observations for the zero-truncated Poisson distribution (143). Using simple calculation, one can show that the parameter λ can be expressed as the root of the following equation,

$$\frac{\lambda}{1 - \exp(-\lambda)} = \frac{\sum_{j \geq 1} j n_j}{\sum_{j \geq 1} n_j}.$$

Once λ is determined, the parameter S can be estimated via a Horvitz–Thompson estimator (47),

$$\frac{\sum_{i \geq 1} n_j}{1 - \exp(-\lambda)}.$$

Another way is to estimate both S and the parameter λ iteratively. One sets an initial value of S and estimates the parameter λ using summary statistics n_j for $j = 0, 1, \ldots$ Based on the estimated parameter λ, the value of S is updated according the Horvitz–Thompson estimator. The process iteratively updates S and λ until they converge. Sanathanan (144, 145) showed that this conditional approach is asymptotically equivalent to the unconditional approach that requires joint maximization of S and the parameter λ.

Note that all the methods typically estimate the total number of species in a local population. In ecology, a local population is the environment where sampling takes place; in high-throughput sequencing, a local population is a prepared sequencing library. In other words, these methods can only infer the number of observable species, which will be observed in the sample as sampling effort goes to infinity. However, in a large number of cases that we considered here, the percent of the genome covered is less than 100% no matter how deep the sample is sequenced due to dropout. The length of the reference genome overestimates the observable length of regions that can be covered by reads. Therefore, if the dropout rate is low, it is reasonable to use the length of the reference genome to approximate the total number of base pairs. Otherwise, it is better to treat the total number of base pairs as unknown.

6.2. Is the Sample Size Sufficient?

The word "sufficient" is ambiguous and depends on objectives. We use library complexity as an example and discuss two scenarios: the reproducibility of a sequencing sample and the capacity of inferring the sequencing library. If sufficiency refers to the reproducibility of a sample, the sample coverage, estimated by $1 - n_1/N$ (see Section 2.2), is then a good, quick, and easy measurement. Low sample coverage means that the probability of obtaining unobserved molecules is high if sequencing continues. This indicates the large randomness of sampling. If a researcher conducts the same sequencing experiment, many molecules observed from the previous experiment may not appear in the current one.

If "sufficiency" means that a sample contains enough information to infer the characteristics of the sequencing library, such as the library complexity curve, the sufficient size depends on the complexity of the library and the relative abundances of molecules in it, which by themselves are hard to estimate. However, the summary statistics contain much useful information and should be used as necessary conditions for a sanity check. In the above, we have discussed sample coverage. Another simple but useful metric, derived from the Good–Toulmin estimator (17), is the alternative sum of summary statistics, $n_1 - n_2 + n_3 - \ldots$, which indicates the expected number of new unique molecules as the sequencing effort is doubled. If the value of this metric is negative, one should be cautious when applying any estimator derived from the mixture assumptions (see Section 2).

In addition, the maximum frequency $j_{\max}$, where $n_j > 0$ for $j \leq j_{\max}$ and $n_{j_{\max}+1} = 0$, is useful to distinguish any outliers that have huge abundances. These outliers possess a large weight on sequencing effort but contain little information about the unseen proportion of the sequencing library. The data with a small value of $j_{\max}$ indicate the instability of high-frequency terms. Therefore, one should put little weight on these terms when constructing an estimator. In particular, Good–Touring frequency estimation of the relative abundance of a read observed r times fails when $n_r = 0$. A smoothing step of n_j is required to incorporate high-frequency terms for estimation (29, 140).

6.3. Is It Safe to Assume that the Observations are Independent?

In high-throughput sequencing it is common to assume that reads are independently generated from a sequencer. However, in terms of genome coverage, the independence of coverage for each

base pair is not obvious. One read can cover multiple sites in the genome, and therefore coverages for nearby sites are related. Note that this is local dependence. When the distance of two sites is greater than the length of a read for single-end sequencing or greater than the length of a fragment for pair-end sequencing, their coverage can be considered independent. According to Chen–Stein method (146, 147), coverage of each base pair in the genome can be approximated by independent Poisson distributions. Unfortunately, the local dependence does not hold for SNPs. Each individual can theoretically identify up to 3 billion SNPs. In contrast, only thousands of individuals are sequenced even for large-scale sequencing projects. This is a typical high-dimensional problem where the number of samples is much smaller than the number of possible SNPs. How to address this issue is the key to improving the estimation of genetic diversity.

6.4. Is Sampling Effort Linear to Observations Obtained?

The total number of observations is calculated as $N = \sum_{j \geq 1} jn_j$. In the application of genome coverage, N is the total number nucleotides from aligned reads. Therefore, strictly speaking, the predicted SAC based on n_j should be interpreted as the number of covered base pairs versus the number of aligned reads. Since there is little difference in the mapping rate between sequencing runs, the number of aligned reads is linear to the number of raw reads, which is the sampling effort.

To ensure the predicted SAC reflects the change of the sequencing effort, the implicit assumption is that the total number N of observations should be approximately linear to the sampling effort. One counterexample is predicting the number of covered base pairs as a function of sequencing effort in WES experiments. Compared with WGS, the coverage of each base pair in the targeted exome regions is calculated based on uniquely aligned reads (148). In other words, if multiple reads are aligned to the same location in the genome, only one read is accounted for in the coverage. Therefore, the predicted curve based on n_j should be interpreted as the number of covered base pairs as a function of the number of uniquely aligned reads, which is not equivalent to the sampling effort. To obtain the SAC, one needs to estimate the duplicate rate as a function of the sampling effort (e.g., 19) and convert the number of uniquely aligned reads to the equivalent sampling effort.

6.5. Is There a Parametric Form that Describes the Population?

Parametric estimation is easy to calculate and interpret but can suffer if the parametric family is incorrect. Nonparametric estimation, in contrast, is robust to misspecification but is typically difficult to calculate. This inherent trade-off leads to the different uses of the two paradigms. For example, in differential expression or peak identification, the ease of interpretation is critical and so parametric estimation is preferred, although techniques such as empirical base smoothing help to improve estimation. We have found that estimation of library complexity, sequencing coverage, and species richness is difficult in the parametric setting. In the case of complexity and coverage estimation, we have found that for accurate estimation, extremely high-dimensional mixtures are required. This lends support to the use of nonparametric estimators, as high-dimensional mixtures are difficult to interpret and properly perform model selection. In the case of species richness, the penalty for using the wrong parametric family can be large (149). Although a multitude of methods are available available, we have found that most nonparametric methods perform better than parametric estimation with a misspecified family. In other words, it is better to use a wrong, but flexible, nonparametric method than it is to use the wrong parametric method.

6.6. Can Additional Sampling be Conducted Identically?

A critical assumption in applying a species sampling model to high-throughput sequencing is that identification of duplicates does not depend on the sequencing depth. When using mapping or *k*-mers to identify duplicates, this assumption holds. But in applications such as microbial sequencing, scRNA-seq with CBs and UMIs, and SNP calling in WGS, these assumptions might not hold. When clustering reads into OTUs, the OTUs can depend on the number of reads available, as more reads will help to distinguish similar species. In scRNA-seq, either UMIs and CBs are clustered or minimum abundance thresholds are put in place to help distinguish dead cells from live cells and to ensure accurate gene expression quantification. This means that a lot of singletons will be excluded from the analysis. It is therefore likely that the true number of singletons is much higher than the estimated number, and the estimation of the sample coverage and library complexity is likely more conservative than the truth. In SNP calling, the power of detecting a SNP depends on the number of reads that cover the base pair. This is particularly salient in SNP calling of tumors, as the SNP frequency can vary wildly (150).

6.7. Evaluation

There are several ways to evaluate the performance of estimators. If both shallow sequencing and deep sequencing data sets are available, one can use the estimator to predict the SAC based on the shallow sequencing data set and extrapolate the result to the size of the deep sequencing data set to evaluate the prediction accuracy. When these data are not available, a convenient way to evaluate performance is subsampling the entire sequencing data set to simulate a shallow sample. Another way is to compare the predicted SAC with the true SAC. Since the accumulation process of the data is usually not available, the best way is to generate a SAC by subsampling the entire data set without replacement. Heck et al. (151) provided a method to calculate the expected number of species versus sampling effort, based on sampling without replacement. If $S_1(t)$ is the number of distinct reads obtained from a subsample of tN reads for $0 \leq t \geq 1$, then we have

$$\hat{\mathrm{E}}[S_1(t)] = \sum_j n_j - \binom{N}{tN}^{-1} \sum_j n_j \binom{N-j}{tN}. \quad 4.$$

The generalized binomial coefficients for noninteger y are

$$\binom{x}{y} = \begin{cases} \frac{\Gamma(x+1)}{\Gamma(y+1)\Gamma(x-y+1)}, & \text{if } x, y \geq 0 \text{ and } x-y \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

In general, the expected number $S_r(t)$ of species represented at least r times versus sampling effort can be estimated as

$$\hat{\mathrm{E}}[S_r(t)] = \sum_j n_j - \binom{N}{tN}^{-1} \sum_j n_j \left[\binom{N-j}{tN-i} \binom{j}{i} \right]. \quad 5.$$

This unbiased estimator has the minimum variance among all unbiased estimators (152). Note that $\hat{\mathrm{E}}[S_r(t)]$ cannot be applied to measure the saturation of the sequencing data, because $\hat{\mathrm{E}}[S_r(t)]$ is undefined for $t > 1$.

For parametric methods, goodness of fit can be used to evaluate the performance of the estimator (153). However, goodness of fit cannot guarantee good behaviors for extrapolation. Engen

(153) provided an example where, although estimators fit the data well, extrapolation for these methods was quite different.

Not all estimators have a closed form for the confidence interval. In practice, one could always use bootstrapping to generate the confidence interval (154, 155). Given a sequencing data set, one can generate many sequencing data sets by sampling reads with replacement and then constructing confidence intervals from the bootstrapped estimates.

7. DISCUSSION

In this review, we have discussed the missing species problem in modern sequencing experiments. We showed that many variations of this problem currently exist in the planning of experiments, allocation of sequencing resources, quality control, estimation of immune repertoire, and genetic diversity. We imagine that many more variations exist that we did not discuss. We hope researchers can become aware of the connections between high-throughput sequencing and the classical species sampling problem, and can leverage methods from capture–recapture statistics to address similar problems in their daily studies.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We would like to thank Mike Waterman for his excellent guidance and advice, Adams River for bringing up the question of metagenomics, Susan Holmes for her helpful discussions and advice on microbial and metagenomic applications, Jie Ren for her wonderful support, and the Smith and Wong labs for their helpful comments.

LITERATURE CITED

1. Pachter L. 2013. *Seq. *Bits of DNA: Reviews and Commentary on Computational Biology by Lior Pachter*, Nov. 23, 2013, accessed Sept. 6, 2018. **https://liorpachter.wordpress.com/seq**
2. Bunge J, Fitzpatrick M. 1993. Estimating the number of species: a review. *J. Am. Stat. Assoc.* 88:364–73
3. Fisher RA, Corbet AS, Williams CB. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 12:42–58
4. Good IJ. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *J. Stat. Comput. Simul.* 66:101–11
5. Johndrow JE, Lum K, Manrique-Vallier D. 2016. Estimating the observable population size from biased samples: a new approach to population estimation with capture heterogeneity. arXiv:1606.02235 [stat.ME]
6. Anscombe FJ. 1950. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* 37:358–82
7. Zipf GK. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin
8. Bulmer MG. 1974. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* 30:101–10
9. Burrell QL, Fenton MR. 1993. Yes, the GIGP really does work—and is workable! *J. Am. Soc. Inform. Sci.* 44:61–69
10. Sichel HS. 1975. On a distribution law for word frequencies. *J. Am. Stat. Assoc.* 70:542–47

11. Norris JL, Pollock KH. 1998. Non-parametric MLE for Poisson species abundance models allowing for heterogeneity between species. *Environ. Ecol. Stat.* 5:391–402
12. Wang JPZ, Lindsay BG. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Stat. Assoc.* 100:942–59
13. Favaro S, Lijoi A, Mena RH, Prünster I. 2009. Bayesian non-parametric inference for species variety with a two-parameter Poisson–Dirichlet process prior. *J. R. Stat. Soc. Ser. B* 71:993–1008
14. Lindsay BG. 1983. The geometry of mixture likelihoods: a general theory. *Ann. Stat.* 11:86–94
15. Wang JP. 2010. Estimating species richness by a Poisson-compound gamma model. *Biometrika* 97:727–40
16. Hansen B, Pitman J. 2000. Prediction rules for exchangeable sequences related to species sampling. *Stat. Probab. Lett.* 46:251–56
17. Good IJ, Toulmin GH. 1956. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43:45–63
18. Burnham KP, Overton WS. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65:625–33
19. Daley T, Smith AD. 2013. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* 10:325–27
20. Deng C, Daley T, Smith A. 2015. Applications of species accumulation curves in large-scale biological data analysis. *Quant. Biol.* 3:135–44
21. Valiant G, Valiant P. 2016. Instance optimal learning of discrete distributions. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*. New York: Assoc. Comput. Mach.
22. Daley T. 2014. *Non-parametric models for large capture-recapture experiments with applications to DNA sequencing*. Ph.D. Thesis, Univ. South. Calif., Los Angeles, CA
23. Zou J, Valiant G, Valiant P, Karczewski K, Chan SO, et al. 2016. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat. Commun.* 7:13293
24. Deng C, Daley T, Calabrese P, Ren J, Smith AD. 2018. Estimating the number of species to attain sufficient representation in a random sample. arXiv:1607.02804 [stat.ME]
25. Raghunathan A, Valiant G, Zou J. 2017. Estimating the unseen from multiple populations. arXiv:1707.03854 [cs.LG]
26. Dumitrascu B, Feng K, Engelhardt BE. 2018. GT-TS: experimental design for maximizing cell type discovery in single-cell data. bioRxiv 386540. **https://doi.org/10.1101/386540**
27. Zipf GK. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard Univ. Press
28. Robbins H. 1964. The empirical Bayes approach to statistical decision problems. *Ann. Math. Stat.* 35:1–20
29. Good IJ. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40:237–64
30. Good IJ. 1979. Studies in the history of probability and statistics. XXXVII: A. M. Turing's statistical work in World War II. *Biometrika* 66:393–96
31. Lladser ME, Gouet R, Reeder J. 2011. Extrapolation of urn models via Poissonization: accurate measurements of the microbial unknown. *PLOS ONE* 6:e21105
32. Lijoi A, Mena RH, Prünster I. 2007. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* 94:769–86
33. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* 349(6250):aab3884
34. Bunge J. 2011. Estimating the number of species with CatchAll. *Pac. Symp. Biocomput.* 2011:121–30
35. Harris B. 1968. Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. *J. Am. Stat. Assoc.* 63:837–47
36. Link WA. 2003. Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* 59:1123–30

37. Holzmann H, Munk A, Zucchini W. 2006. On identifiability in capture-recapture models. *Biometrics* 62:934–36
38. Mao CX, Lindsay BG. 2007. Estimating the number of classes. *Ann. Stat.* 35(2):917–30
39. Daley T, Smith AD. 2018. Better lower bounds: improved non-parametric moment-based species estimation for large experiments. arXiv:1605.03294 [stat.ME]
40. Lindsay BG. 1995. Mixture models: theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Vol. 5, pp. i–163. Hayward, CA: Inst. Math. Stat.
41. Wang JP, Lindsay BG. 2008. An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Stat. Methodol.* 5:30–45
42. Wang JPZ, Lindsay BG. 2005. A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Stat. Assoc.* 100:942–59
43. Kaplinsky J, Arnaout R. 2016. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat. Commun.* 7:11881
44. Valiant G, Valiant P. 2013. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. **https://papers.nips.cc/paper/5170-estimating-the-unseen-improved-estimators-for-entropy-and-other-properties**
45. Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 11:265–70
46. Burnham KP, Overton WS. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65:625–33
47. Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47:663–85
48. Chao A, Lee SM. 1992. Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* 87:210–17
49. Chao A, Bunge J. 2002. Estimating the number of species in a stochastic abundance model. *Biometrics* 58:531–39
50. Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, et al. 2014. Diversity and clonal selection in the human T-cell repertoire. *PNAS* 111:13139–44
51. Willis A, Bunge J. 2015. Estimating diversity via frequency ratios. *Biometrics* 71:1042–49
52. Böhning D, Rocchetti I, Alfó M, Holling H. 2016. A flexible ratio regression approach for zero-truncated capture–recapture counts. *Biometrics* 72:697–706
53. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, et al. 2009. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* 114:4099–107
54. Laydon DJ, Melamed A, Sim A, Gillet NA, Sim K, et al. 2014. Quantification of HTLV-1 clonality and TCR diversity. *PLOS Comput. Biol.* 10:e1003646
55. Willis A. 2016. Extrapolating abundance curves has no predictive power for estimating microbial biodiversity. *PNAS* 113:E5096
56. Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–10
57. Elhanati Y, Murugan A, Callan CG, Mora T, Walczak AM. 2014. Quantifying selection in immune receptor repertoires. *PNAS* 111:9875–80
58. Mangul S, Yang HT, Strauli N, Gruhl F, Porath HT, et al. 2018. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.* 19:36
59. Tian L, Fire AZ, Boyd SD, Olshen RA. 2018. Clonality: point estimation. *Ann. Appl. Stat.* 63(2):522–30
60. Venturi V, Kedzierska K, Turner SJ, Doherty PC, Davenport MP. 2007. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J. Immunol. Methods* 321:182–95
61. Chao A, Wang Y, Jost L. 2013. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods Ecol. Evol.* 4:1091–100
62. Chao A, Shen TJ. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* 10:429–43
63. Jost L. 2006. Entropy and diversity. *Oikos* 113:363–75

64. Zhang Z, Zhou J. 2010. Re-parameterization of multinomial distributions and diversity indices. *J. Stat. Plan. Inference* 140:1731–38
65. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, et al. 2011. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9:72–74
66. Sahlén P, Abdullayev I, Ramsköld D, Matskova L, Rilakovic N, et al. 2015. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol.* 16:156
67. Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard JM, Poinar HN. 2014. Ancient whole genome enrichment using baits built from modern DNA. *Mol. Biol. Evol.* 31:1292–94
68. Sharon D, Tilgner H, Grubert F, Snyder M. 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* 31:1009–14
69. Gamba C, Hanghøj K, Gaunitz C, Alfarhan AH, Alquraishi SA, et al. 2016. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. *Mol. Ecol. Resourc.* 16:459–69
70. Chandra T, Kirschner K, Thuret JY, Pope BD, Ryba T, et al. 2012. Independence of repressive histone marks and chromatin compaction during senescent heterochromatic layer formation. *Mol. Cell* 47(2):203–14
71. de Almeida CR, Stadhouders R, de Bruijn MJ, Bergen IM, Thongjuea S, et al. 2011. The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. *Immunity*. 35(4):501–13
72. Qu J, Hodges E, Molaro A, Gagneux P, Dean MD, et al. 2018. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Res.* 28(2):145–58
73. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15:121–32
74. Daley T, Smith AD. 2014. Modeling genome coverage in single-cell sequencing. *Bioinformatics* 30:3159–65
75. Zong C, Lu S, Chapman AR, Xie XS. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338:1622–26
76. Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231–39
77. Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philos. Trans. R. Soc. B* 345:101–18
78. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, et al. 2014. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32:1053–58
79. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, et al. 2018. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* 174:1015–30
80. Ekblom R, Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* 7:1026–42
81. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, et al. 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLOS Comput. Biol.* 9:e1003326
82. Illumina. 2014. *Estimating sequencing coverage*. Tech. Note 770-2011-022, Illumina, San Diego, CA. **https://www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf**
83. Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *PNAS* 74:5463–67
84. Staden R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* 8:3673–94
85. Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72
86. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* 26:1293–300
87. Hooper SD, Dalevi D, Pati A, Mavromatis K, Ivanova NN, Kyrpides NC. 2010. Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics* 26:295–301

88. Miller CA, Hampton O, Coarfa C, Milosavljevic A. 2011. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLOS ONE* 6:e16327
89. 1,000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526:68–74
90. Woodworth MB, Girskis KM, Walsh CA. 2017. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* 18:230–44
91. Colwell RK, Chao A, Gotelli NJ, Lin SY, Mao CX, et al. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* 5:3–21
92. Ionita-Laza I, Lange C, Laird NM. 2009. Estimating the number of unseen variants in the human genome. *PNAS* 106:5008–13
93. Gravel S. 2014. Predicting discovery rates of genomic features. *Genetics* 197:601–10
94. Efron B, Thisted R. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* 63:435–47
95. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. 2011. Demographic history and rare allele sharing among human populations. *PNAS* 108:11983–88
96. Burnham KP, Overton WS. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65:625–33
97. Xu C. 2018. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* 16:15–24
98. Schwartz RS. 2003. Diversity of the immune repertoire and immunoregulation. *N. Engl. J. Med.* 348:1017–26
99. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. 2015. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. B* 370:20140243
100. Mora T, Walczak A. 2016. Quantifying lymphocyte receptor diversity. arXiv:1604.00487 [q-bio.PE]
101. den Braber I, Mugwagwa T, Vrisekoop N, Westera L, Mögling R, et al. 2012. Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* 36:288–97
102. Mora T, Walczak AM, Bialek W, Callan CG. 2010. Maximum entropy models for antibody diversity. *PNAS* 107:5405–10
103. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, et al. 2014. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* 192:2689–98
104. Gibson KL, Wu YC, Barnett Y, Duggan O, Vaughan R, et al. 2009. B-cell diversity decreases in old age and is correlated with poor health status. *Aging Cell* 8:18–25
105. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, et al. 2008. Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. *Nature* 453:667–71
106. Qi Q, Cavanagh MM, Le Saux S, Wagar LE, Mackey S, et al. 2016. Defective T memory cell differentiation after varicella zoster vaccination in older individuals. *PLOS Pathog.* 12:e1005892
107. Li B, Li T, Pignon JC, Wang B, Wang J, et al. 2016. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* 48:725–32
108. Perline R. 2005. Strong, weak and false inverse power laws. *Stat. Sci.* 20:68–88
109. White EP, Enquist BJ, Green JL. 2008. On estimating the exponent of power-law frequency distributions. *Ecology* 89:905–12
110. Clauset A, Shalizi CR, Newman ME. 2009. Power-law distributions in empirical data. *SIAM Rev.* 51:661–703
111. Warren RL, Freeman JD, Zeng T, Choe G, Munro S, et al. 2011. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Res.* 21:790–97
112. Chiarucci A, Di Biase RM, Fattorini L, Marcheselli M, Pisani C. 2018. Joining the incompatible: exploiting purposive lists for the sample-based estimation of species richness. *Ann. Appl. Stat.* 12:1679–99
113. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6:377–82

114. Menon V. 2017. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Brief. Funct. Genom.* 17:240–45
115. Zamanighomi M, Lin Z, Daley T, Chen X, Duren Z, et al. 2018. Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* 9:2410
116. Trapnell C. 2015. Defining cell types and states with single-cell genomics. *Genome Res.* 25:1491–98
117. Islam S, Kjällquist U, Moliner A, Zajac P, Fan JB, et al. 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* 21(7):1160–67
118. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202–14
119. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161:1187–201
120. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, et al. 2017. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049
121. Taylor K, Watson L, Frenz L, Greiner D, Lebofsky R, et al. 2017. *A scalable high-throughput method for RNA-seq analysis of thousands of single cells*. White Pap. 1070-2016-013, Illumina, San Diego, CA. **https://jp.illumina.com/content/dam/illumina-marketing/documents/products/flyers/ddseq-single-cell-poster-handout-single-cell-poster-handout-web.pdf**
122. Lindström NO, Brandine GDS, Tran T, Ransick A, Suh G, et al. 2018. Progressive recruitment of mesenchymal progenitors reveals a time-dependent process of cell fate acquisition in mouse and human nephrogenesis. *Dev. Cell* 45:651–60
123. Hicks SC, Townes FW, Teng M, Irizarry RA. 2017. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*. 19(4):562–78
124. Kharchenko PV, Silberstein L, Scadden DT. 2014. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11:740–42
125. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. 2018. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* 9:284
126. Chen K, Pachter L. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLOS Comput. Biol.* 1:106–12
127. Human Microbiome Proj. Consort. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486:207–14
128. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, et al. 2017. Strains, functions and dynamics in the expanded human microbiome project. *Nature* 550:61–66
129. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, et al. 2015. Structure and function of the global ocean microbiome. *Science* 348:1261359
130. Bunge J, Willis A, Walsh F. 2014. Estimating the number of species in microbial diversity studies. *Annu. Rev. Stat. Appl.* 1:427–45
131. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13:581–83
132. Ren B, Bacallado S, Favaro S, Holmes S, Trippa L. 2017. Bayesian nonparametric ordination for the analysis of microbial communities. *J. Am. Stat. Assoc.* 112:1430–42
133. Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, et al. 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. bioRxiv 266700. **https://doi.org/10.1101/266700**
134. Edgar R. 2017. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ* 5:e3889
135. Rodriguez-R LM, Konstantinidis KT. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30:629–35
136. Pevzner PA, Tang H, Waterman MS. 2001. An Eulerian path approach to dna fragment assembly. *PNAS* 98:9748–53
137. Marais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27:764–70
138. Wang JP. 2011. Species: an R package for species richness estimation. *J. Stat. Softw.* 40:1–15

139. Chao A, Chiu CH. 2016. Nonparametric estimation and comparison of species richness. *eLS*. **https://doi.org/10.1002/9780470015902.a0026329**
140. Gale WA, Sampson G. 1995. Good-Turing frequency estimation without tears. *J. Quant. Linguist.* 2:217–37
141. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–40
142. Colwell RK, Elsensohn JE. 2014. EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography* 37:609–13
143. Cohen AC. 1960. Estimating the parameter in a conditional poisson distribution. *Biometrics* 16:203–11
144. Sanathanan L. 1972. Estimating the size of a multinomial population. *Ann. Math. Stat.* 43:142–52
145. Sanathanan L. 1977. Estimating the size of a truncated sample. *J. Am. Stat. Assoc.* 72:669–72
146. Chen LHY. 1975. Poisson approximation for dependent trials. *Ann. Probab.* 3:534–45
147. Arratia R, Goldstein L, Gordon L. 1989. Two moments suffice for Poisson approximations: the Chen–Stein method. *Ann. Probab.* 17:9–25
148. Bao R, Huang L, Andrade J, Tan W, Kibbe WA, et al. 2014. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform.* 13:67–82
149. Hong SH, Bunge J, Jeon SO, Epstein SS. 2006. Predicting microbial species richness. *PNAS* 103:117–22
150. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, et al. 2012. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat. Commun.* 3:811
151. Heck KL, van Belle G, Simberloff D. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology* 56:1459–61
152. Xuan Mao C, Colwell RK, Chang J. 2005. Estimating the species accumulation curve using mixtures. *Biometrics* 61:433–41
153. Engen S. 1978. *Stochastic Abundance Models*. London: Chapman and Hall
154. Efron B, Tibshirani RJ. 1994. *An Introduction to the Bootstrap*. London: Chapman and Hall
155. Kuhnert R, del Rio Vilas VJ, Gallagher J, Böhning D. 2008. A bagging-based correction for the mixture model estimator of population size. *Biom. J.* 50:993–1005