Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity

Luis M. Rodriguez-R,[a] Santosh Gunturu,[c] James M. Tiedje,[c,d,e] James R. Cole,[c,e] Konstantinos T. Konstantinidis[a,b]

[a]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA
[b]School of the Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA
[c]Center for Microbial Ecology, Michigan State University, East Lansing, Michigan, USA
[d]Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, Michigan, USA
[e]Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, Michigan, USA

ABSTRACT Estimations of microbial community diversity based on metagenomic data sets are affected, often to an unknown degree, by biases derived from insufficient coverage and reference database-dependent estimations of diversity. For instance, the completeness of reference databases cannot be generally estimated since it depends on the extant diversity sampled to date, which, with the exception of a few habitats such as the human gut, remains severely undersampled. Further, estimation of the degree of coverage of a microbial community by a metagenomic data set is prohibitively time-consuming for large data sets, and coverage values may not be directly comparable between data sets obtained with different sequencing technologies. Here, we extend Nonpareil, a database-independent tool for the estimation of coverage in metagenomic data sets, to a high-performance computing implementation that scales up to hundreds of cores and includes, in addition, a *k*-mer-based estimation as sensitive as the original alignment-based version but about three hundred times as fast. Further, we propose a metric of sequence diversity (N_d) derived directly from Nonpareil curves that correlates well with alpha diversity assessed by traditional metrics. We use this metric in different experiments demonstrating the correlation with the Shannon index estimated on 16S rRNA gene profiles and show that N_d additionally reveals seasonal patterns in marine samples that are not captured by the Shannon index and more precise rankings of the magnitude of diversity of microbial communities in different habitats. Therefore, the new version of Nonpareil, called Nonpareil 3, advances the toolbox for metagenomic analyses of microbiomes.

IMPORTANCE Estimation of the coverage provided by a metagenomic data set, i.e., what fraction of the microbial community was sampled by DNA sequencing, represents an essential first step of every culture-independent genomic study that aims to robustly assess the sequence diversity present in a sample. However, estimation of coverage remains elusive because of several technical limitations associated with high computational requirements and limiting statistical approaches to quantify diversity. Here we described Nonpareil 3, a new bioinformatics algorithm that circumvents several of these limitations and thus can facilitate culture-independent studies in clinical or environmental settings, independent of the sequencing platform employed. In addition, we present a new metric of sequence diversity based on rarefied coverage and demonstrate its use in communities from diverse ecosystems.

KEYWORDS bioinformatics, coverage, metagenomics, microbial ecology

The exploration of microbial diversity in natural and engineered environments has been revolutionized by the use of metagenomics. However, the power of both descriptive and comparative metagenomic analyses is strongly deterred by low cover-

Received 22 March 2018 **Accepted** 23 March 2018 **Published** 10 April 2018

Citation Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. mSystems 3:e00039-18. https://doi.org/10.1128/mSystems.00039-18.

Editor Anthony Fodor, University of North Carolina at Charlotte

Address correspondence to Konstantinos T. Konstantinidis, kostas@ce.gatech.edu.

L.M.R. and S.G. contributed equally to this work.

Nonpareil 3 estimates coverage and sequence diversity from metagenomes without reference databases

age, defined as the fraction of the DNA space covered by sequencing (1). To date, most metagenomic studies assess the level of coverage only indirectly or not at all, mainly owing to the difficulty of directly measuring the unseen fraction of a community. For instance, many studies have assessed coverage by extracting 16S rRNA gene-containing reads from amplicon or shotgun metagenomes, clustering these sequences in operational taxonomic units (OTUs) on the basis of their best matches in reference databases or by *de novo* clustering, and counting the discovery of new OTUs with an increasing number of available sequences. The degree to which this approach represents the real diversity of the sample remains essentially unknown because it depends on the unbiased recovery of 16S rRNA gene-containing reads from the metagenome, the comprehensiveness of the reference database in representing the natural diversity in the sample, the extent of undersampling of the diversity, and the fact that identical 16S rRNA gene sequences may represent distinct species because of the high degree of sequence conservation of the rRNA genes (2).

We have recently presented a method to assess the level of coverage provided by metagenomic data sets by a redundancy-based approach, Nonpareil (3). Briefly, Nonpareil estimates the read redundancy in a metagenomic data set by using ungapped alignments of a subset of sequencing reads against the entire metagenome and then applies the Turing-Good estimator principle to approximate the abundance-weighted coverage of the metagenome. Abundance weighting means that the coverage estimate represents the fraction of the total DNA extracted from a sample (and by extension, the fraction of the organisms, not species) that is represented in a set of metagenomic sequence data, not necessarily how many different species are represented by the extracted versus the sequenced DNA. For instance, the nonsequenced fraction may represent a higher number of species even in cases where the coverage is >50% if the community is characterized by high species evenness, which is not uncommon for soil and sediment habitats. Nonpareil is database independent because it is based on intrinsic characteristics of the data set and, given sufficient data (typically, a small fraction of most available metagenomes), can produce robust estimations of coverage regardless of the sequencing effort applied.

Two previously described techniques for the estimation of coverage in metagenomes exist. The first one, COVER (4), assumes reference database completeness and a high degree of taxonomic conservation in 16S rRNA gene copy number and genome size, assumptions often violated by available genome sequences (5). Accordingly, Nonpareil is substantially more accurate than COVER for most complex communities as previously shown (3). The second one is a parametric approach currently not implemented in any available standalone tool (6) that models occupancy as a Poisson process using a gamma approximation and requires estimates of abundance and genome size joint distributions. Other approaches related to the problem of coverage exist but offer only indirect measurements such as the maximum expected contig size (7) or the coverage of a target minimum-abundance organism (8, 9) from which only an implementation for viral communities exists, i.e., MetLab (10). Therefore, Nonpareil is advantageous compared to previously described approaches and bioinformatic tools in that it is database independent, directly estimates coverage, and is fully automated, allowing the end user to simply input a metagenomic data set for coverage estimation. In addition, Nonpareil subsamples the estimated coverage and fits the rarefied curve to a sigmoidal function in order to predict the sequencing effort necessary to reach any target coverage (3). Using Nonpareil, we have shown that coverage affects not only the completeness of descriptive metagenome-based profiling but also the accuracy of comparative abundance analyses of features such as species or genes, highlighting the importance of coverage estimations for both descriptive and comparative metagenomics (1).

While we have previously demonstrated that Nonpareil accurately estimates the level of coverage in less complex metagenomic data sets, the estimation remained prohibitively expensive for data sets comprising several billion base pairs. Here, we present a new version of Nonpareil, Nonpareil 3, that effectively distributes the estimation across processors and computing nodes by using high-performance comput-

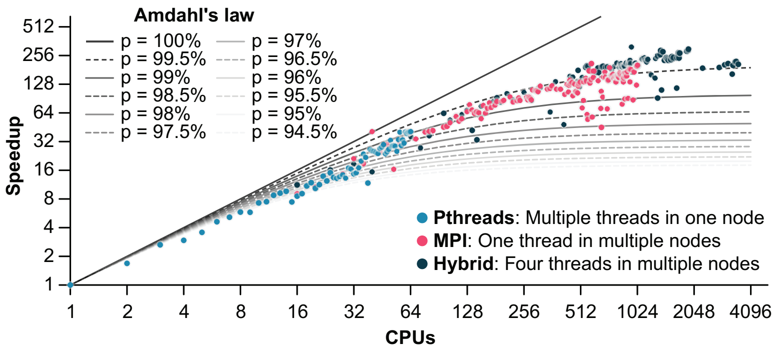

FIG 1 Speedup per number of processors in Nonpareil. Nonpareil estimations of the LL_1007B data set (3) were performed with alignment kernel and default parameters in multiple processors of a node (light blue), a single processor of multiple nodes (pink), and four processors of multiple nodes (dark green). The base time (one processor of one node) was 82.98 h.

ing capabilities now often available to laboratories working with metagenomic data. Nonpareil 3 also includes a new algorithm providing an alternative estimation of coverage based on *k*-mer redundancy, as opposed to ungapped alignment in the original version, with comparable accuracy but 2 orders of magnitude faster computation. This version allows end users to run Nonpareil analysis of relatively large metagenomic data sets on personal computers as well. Moreover, we previously showcased the qualitative sequence diversity rankings derived from Nonpareil curves (1, 3). Here, we quantitatively assessed the level of sequence diversity derived from Nonpareil curves, compared these estimations with other diversity indices, and present typical ranges for communities from different environments, allowing for quantitative ranking of the communities on the basis of their sequence complexity.

RESULTS

Reducing run time for large metagenomes. Nonpareil 3 can use parallelization across nodes and threads by using Message Passing Interface (MPI) and pthreads, respectively. We executed Nonpareil with the original alignment kernel by using both parallelization methods, independently and in combination, on a 2.3-Gbp test data set in order to evaluate the speedup. These optimizations resulted in up to 500 times faster computation of coverage. Compared to Amdahl's law, Nonpareil 3 speedups corresponded to around 99.5% parallelization with MPI alone and around 99.8% parallelization with MPI and four threads, while the observed speedup for multithreads in one node was essentially linear (Fig. 1). In addition, the read redundancy can now be estimated by using perfect matches of one *k*-mer per query read corrected by sequencing error estimation, instead of the complete ungapped alignment. This implementation results in similar coverage and diversity estimates, as well as highly correlated projections of sequencing effort, but reduces the computational time by about 300 times (Table 1; see Fig. S1 in the supplemental material). We recommend using $k = 24$ (the default value in Nonpareil), the smaller value producing stable estimates (Fig. S2). The runtime of Nonpareil with alignment kernel is approximately proportional to $N \times Q \times L^2$, where N is the total number of sequencing reads, Q is the number of query sequencing reads searched against the complete metagenomic data set, and L is the length of a read. In contrast, the complexity of Nonpareil with *k*-mer kernel is simply $N \times L$, i.e., directly proportional to the total data set size (Fig. S3). As the runtime for the *k*-mer kernel is independent of the read length, Nonpareil 3 is directly compatible with long-read data, although current technologies (e.g., PacBio or MinION) exhibit error rates that are higher than optimal for the current Nonpareil implementation (i.e., $<5\%$ expected sequencing error). Thus, additional optimizations will be necessary to allow reads with higher sequencing error to be analyzed by Nonpareil. In addition, since the runtime is not proportional to Q, it is practical to use a larger value for Q (default of

TABLE 1 Kernel comparison of Nonpareil estimates for publicly available data sets[a]

				CPU time (min)		% coverage		Required effort (Gbp)	
Sample	Identifier	Reference	Size (Gbp)	A	K	A	K	A	K
Posterior fornix	SRS063417	25	0.01	15.7	0.08	89	84	0.062	0.070
Stool sample	SRS015540	25	0.32	438	0.85	81	71	2.62	5.55
Tongue	SRS055495	25	0.22	286	0.68	71	61	3.22	6.08
LL 2011	SRR948155	3	2.95	4,397	16.5	84	79	11.7	24.1
LL 2009A	SRR096386	26	1.17	1,444	6.40	68	64	20.5	24.8
LL 2009B	SRR096387	26	1.12	1,463	5.75	70	64	14.3	20.0
Iowa soil	JGI 402461	NA[b]	14.6	22,806[c]	49.0	56	48	662	1,051

[a] The two kernels (A, alignment; K, *k*-mer) were compared in terms of CPU time, estimated coverage, and projected required sequencing effort to reach 95% coverage of samples varying in complexity, including HMP (posterior fornix, tongue, stool sample), freshwater (Lake Lanier [LL]), and soil (Iowa continuous cornfield).

[b] NA, not available.

[c] CPU time was estimated for Iowa soil and observed in all other cases.

10,000 instead of 1,000 in the alignment kernel), increasing the precision of Nonpareil over that of the original version.

Sequencing error correction. In order to reduce the impact of sequencing error on Nonpareil with *k*-mer kernel, we implemented a redundancy correction based on the sequencing read quality scores. Briefly, we estimated the fraction of selected *k*-mers that are expected to include at least one sequencing error from the quality scores and removed that fraction from the list of *k*-mers without matches, assuming that the introduced errors would most frequently result in unobserved variants (see Materials and Methods for details). To test this correction, low- and high-coverage simulated metagenome-like data sets of 101-bp reads (507,813 and 7,093,697 reads, respectively) were generated *in silico* from 30 bacterial and archaeal genomes by a previously described method (3) (Table S1; data sets are available at http://enve-omics.ce.gatech.edu/data/nonpareil). For each test, 10,000 reads were randomly selected from the simulated data sets and modified with substitution errors at a constant error probability per base. The 10,000 reads were then used in Nonpareil with *k*-mer kernel as query sequences with error correction enabled and disabled (Table S2; Fig. S4). In general, error rates affected the estimation of both coverage and required sequencing effort when error correction was disabled but not when it was enabled, indicating that the correction effectively removes the effect of sequencing errors when the error probability estimation is accurate.

Nonpareil N_d. The Nonpareil index of sequence diversity (N_d), described in Materials and Methods, is expressed in units of the natural logarithm of base pairs and summarizes the community diversity in sequence space, i.e., how redundant (or conversely unique) the sequences of a data set are among themselves. This metric depends on the joint distribution of genome size and abundance, as well as intragenome gene duplication. Therefore, given a small variation in genome size and a small impact of genomic duplications, e.g., for prokaryote-only communities, N_d can be used as a database-independent metric of alpha diversity. Since the shapes of the Nonpareil curves from replicates and subsamples closely resemble each other regardless of coverage (3), we propose N_d as a coverage-independent measurement of the diversity of the sampled community (Fig. S5).

We compared N_d and Shannon diversity index values in natural units (H') from 16S rRNA gene OTU tables in different collections of metagenomic data sets, including: set I, metagenomes from different biomes including six mock data sets; set II, metagenomes and 16S rRNA gene amplicons from human microbiomes of different body sites; and set III, a collection of marine metagenomes from different sampling sites. First, we observed a high correlation (Pearson's $r = 0.804$ [$P < 10^{-15}$]; analysis of variance [ANOVA], 64.6% variance explained [$P < 10^{-15}$]) between N_d and H' values from diverse environments (set I), as well as a monotonic trend for interquartile ranges (IQRs) between different biomes (Fig. 2). This correlation was maintained in the subset of mock samples alone ($r = 0.87$ [$P = 0.023$]), and the residuals were only slightly affected

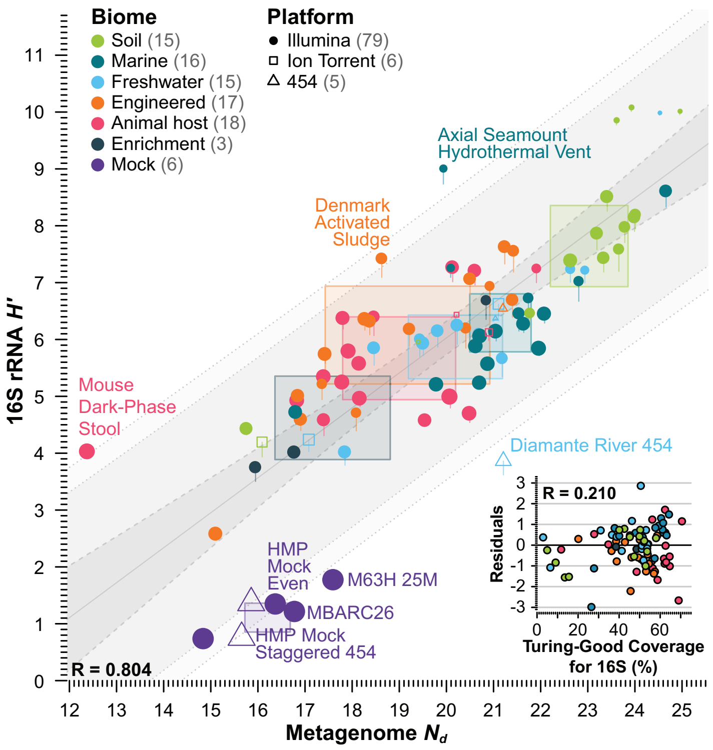

FIG 2 Comparison of Nonpareil N_d sequence diversity and 16S rRNA gene OTU Shannon H' taxonomic diversity indices on 90 metagenomes. Each data point represents the estimates on N_d (*x* axis) and H' (*y* axis). The *y*-axis value of each point indicates the Bayesian analysis-corrected Shannon index, and the line extending from the low part of each data point represents the exact observed (maximum-likelihood) Shannon index. The color of each point indicates the type of biome of each data set, the shape indicates the sequencing platform, and the size indicates the estimated coverage of the 16S rRNA gene profile (Turing-Good estimate). For each biome, the IQR of both estimates is represented as semitransparent rectangles. The least-squares linear correlation model is represented in gray, including the central estimate (solid line), the 95% confidence interval (dashed-line band), and the 80 and 95% prediction intervals (dotted-line bands). Labeled data sets fell outside the 80% prediction interval. The inset shows the residuals from the linear model against the Turing-Good estimate of 16S rRNA gene coverage.

by sequencing technology (ANOVA after including N_d, 3.7% variance [$P = 0.008$]). However, most outliers had low Turing-Good coverage estimates for the 16S rRNA gene OTU count profiles (labeled points in Fig. 2). Indeed, we observed a significant effect of the 16S rRNA gene Turing-Good coverage estimates on the residuals (ANOVA after N_d, excluding mock samples, 5.7% variance [$P = 10^{-4}$]; ANOVA after N_d, including mock samples assuming complete coverage, 14.7% variance [$P = 10^{-11}$]) (Fig. 2, inset). We did not observe a significant effect of the Nonpareil-estimated coverage on the model residuals (ANOVA after N_d, 0.008% variance [$P = 0.9$]), indicating that a significant component of the difference between the two estimates of diversity was due to insufficient data to robustly estimate H' but not N_d. We were able to differentiate between the effect of coverage on both estimates independently because the correlation between 16S rRNA gene-derived Turing-Good coverage estimates and metagenome-derived Nonpareil coverage estimates was partially lost by subsampling of most metagenomic data sets in the estimation of N_d but not in the estimation of H' ($r = 0.64$). (Note that H' was estimated on the basis of previously constructed OTU tables on the complete data sets available [11], not subsamples as in the case of N_d; see Materials and Methods for details.)

Next, we evaluated paired metagenomic and 16S rRNA gene amplicon samples from the Human Microbiome Project (HMP) (set II) and observed a similarly high correlation between median N_d and median H' values per body site ($r = 0.93$), as well as between N_d and H' values per data set ($r = 0.84$; Fig. S6).

Finally, we explored a collection of marine data sets (set III) and compared N_d and H' diversity indices against available metadata by using ANOVA with alpha = 0.01. After controlling for the variance introduced by the source project (as the first independent variable in ANOVAs), we evaluated the effects of different variables on diversity variation (see Materials and Methods). For both measurements of diversity (N_d and H'), we observed significant effects of size fraction (H' variance explained, 16%; N_d variance explained, 25% [$P < 10^{-11}$]) and latitude (H' variance explained, 6.7%; N_d variance explained, 1.5% [$P < 0.01$]), similar to previously reported results for estimated richness as the measure of alpha diversity (12). Geographic location was also found to significantly affect the variance of N_d (12% variance, $P = 10^{-8}$) and much more weakly so the variance of H' (4.3% variance, $P = 0.02$). Moreover, N_d captured a pattern of increased diversity toward the winter (wintriness; explained variance, 2.3% [$P = 0.0011$]) not observed with H' (explained variance, 0.14% [$P = 0.46$]).

DISCUSSION

Estimations of coverage from sequencing data have been a problem studied since the onset of DNA sequencing techniques (reviewed in reference 13). However, most efforts have been directed at estimation of the coverage of a single genome such as those proposed early on by Lander and Waterman (14), as well as more refined models later proposed by Wendl and collaborators (8, 9, 15). Few attempts have been made to extend some of these concepts to entire communities assuming joint distributions for abundance, genome size, and other relevant genomic features (6, 7). However, such attempts have only indirectly addressed the problem of coverage, ultimately targeting other characteristics such as the maximum expected contig length. Therefore, these approaches are not directly comparable to the Nonpareil estimate of coverage. Intrinsic characteristics of random samplings (such as metagenomic data sets) can also be leveraged to determine the level of coverage by using the principle of the Turing-Good estimator (16, 17). This principle has been applied to the estimation of diversity in OTU profiles (18, 19), and similar approaches have been explored for the extrapolation of any count statistics derived from sequencing data (20). Nonpareil uniquely uses this principle to directly estimate the abundance-weighted average coverage of a metagenomic data set based on the degree of overlap of individual metagenomic reads (3).

We have implemented algorithmic and computational improvements in Nonpareil 3 that now make it feasible to process large metagenomic data sets (e.g., tens to hundreds of gigabase pairs) in minutes to hours, even with modest computational resources and without compromising accuracy. Moreover, in addition to the estimation of abundance-weighted average coverage and the projection to estimate required sequencing efforts, Nonpareil 3 includes an estimation of sequence diversity, N_d. We demonstrate that N_d correlates well with alpha diversity derived from 16S rRNA gene OTU profiles but can also capture patterns in diversity not observed with 16S rRNA gene profiles, likely because of the increased taxonomic resolution attainable by whole-genome analyses. For example, seasonal diversity patterns observed with N_d in marine metagenomic data sets were not captured by 16S rRNA gene OTU diversity measured by Shannon H'. Therefore, Nonpareil advances the molecular toolbox for environmental surveys, providing estimations of coverage and diversity independent of databases and robust to various levels of sequencing effort applied, required sequencing effort for complete coverage, and sequence diversity, and scales well with large metagenomic data sets. Applying Nonpareil to metagenomic data sets from different biomes allowed us to quantify sequence diversity and its typical range for different environments. Unsurprisingly, the most diverse communities were those in soil, with an N_d IQR of 22 to 24. Marine communities (open ocean) followed soil communities at about 2 units lower, with an N_d IQR of 20.5 to 21.8. Because of the logarithmic nature of N_d, this corresponds to a sequence diversity about seven times lower in marine than in soil communities. We observed wider and largely overlapping ranges of N_d in communities from freshwater (IQR, 19.5 to 21.1), engineered systems (IQR, 17.8 to 20.7), and animal hosts (IQR, 18.1 to 20.3), all of which are about 2 to 7 times less diverse than those in the open ocean and about 15 to 45 times less diverse than those in soil. Importantly, these differences in sequence diversity translate to differences of several orders of magnitude in required sequencing effort in order to achieve a similar level of coverage (3). Human-associated communities ($n = 8$) did not differ significantly from nonhuman animal-associated samples (N_d IQR, 18.2 to 20.1 versus 17.6 to 20.3), including 4 mouse samples, 2 cow samples, 2 pig samples, 1 chicken sample, and 1 salmon sample ($n = 10$). These results should represent useful reference points for designing future metagenomic studies and achieving the level of coverage that would be desirable for each project and its research objective(s).

MATERIALS AND METHODS

Implementation. The processing of a sample in Nonpareil is divided into two main steps. The first is redundancy estimation, where sequences are compared and the estimated redundancy is subsampled at different values of sequencing effort. The large number of pairwise alignments makes this step the most resource intensive. This task is now distributed across multiple nodes using MPI and processors using C++ pthreads. Further, Nonpareil 3 offers a *k*-mer-based method for redundancy estimation (see below) that accelerates this step by using short fragments of the sequencing reads with no errors allowed. Finally, redundancies are subsampled in multiple threads. In Nonpareil 1, subsamples were estimated linearly, resulting in sparse values toward the left side of the Nonpareil curve. Although this strategy is still available, the default in Nonpareil 3 is logarithmic subsampling; sample size is iteratively multiplied by a density factor (default 0.7) until only two reads remain.

The second step is estimation of the abundance-weighted average coverage at different sequencing efforts (Nonpareil curves), fitting to a sigmoidal model (projection), and graphical representation (e.g., Fig. S1). Note that this step relies on the assumption of independence of events between sequencing reads; therefore, Nonpareil should be applied on single reads, one of the sister reads in paired-end reads, or merged paired-end reads (3). This step has modest resource requirements and is implemented in the Nonpareil R package. In Nonpareil 3, we have streamlined this analysis by using headers in the redundancy output files and included an estimation of the sequence diversity derived from the fitted model (see below). All of the experiments in this report were executed with Nonpareil v3.3.

***k*-mer kernel.** To accelerate read-to-read comparisons, Nonpareil 3 now implements a *k*-mer-based comparison kernel. Briefly, query *k*-mers are derived from a randomly selected subset of the metagenomic data set reads, with one target *k*-mer selected from the 5′ end of each read. Determination of the number of times each target *k*-mer (or its reverse complement) is found at any position of any sequence read in the complete data set can be performed in time proportional to the size of the metagenomic data set and is independent of the number of target *k*-mers. Note that this test is unable to detect if any of the last $k - 1$ bases in a sequencing read cover the target *k*-mer, so for each metagenomic read of length L, only $L - k + 1$ positions can be tested for matches, reducing the effective size of the metagenomic data set scanned with respect to the complete alignment kernel. Nonpareil curves (R package) account for the effect of this difference between sequencing effort and effective size, and its effect is corrected.

Because the target *k*-mers are derived from the metagenomic sequence reads, some of the *k*-mers will contain sequence errors, but if *k* is large enough, these *k*-mers will likely have zero coverage. Although it is not possible to determine which of the zero-coverage target *k*-mers contain errors, we utilized the sequencing quality (Q) scores from each target *k*-mer's parent sequencing read to estimate E, the expected number of *k*-mers with errors. To correct for these errors, we reduced the target set by removing E zero-coverage target *k*-mers. The coverage counts for the remaining target *k*-mers, along with the reduced effective metagenomic data set size, are passed through the remaining original Nonpareil steps to estimate coverage, required sequencing effort, and sequence diversity.

***k*-mer kernel comparison to the original alignment-based kernel.** To compare the results obtained with the traditional alignment kernel and the novel *k*-mer kernel, we executed both analyses in seven metagenomic data sets with different degrees of diversity (Table 1; Fig. S1). Metagenomic data sets were processed by using SolexaQA (21) with a maximum expected error of 1% and a minimum length of 50 bp, and adapter contamination was clipped by using Scythe (https://github.com/vsbuffalo/scythe). For paired-end samples, only the forward reads were used. Short Read Archive (SRA) identifiers are provided in Table 1 for all of the data sets except Iowa continuous cornfield soil. For the latter, seven lanes from one run of Illumina HiSeq were retrieved from the JGI Genome Portal (http://genome.jgi.doe.gov) on 21 July 2016 from project 402461.

Nonpareil results with *k*-mer kernel were obtained by using a 27-in. iMac with 8 gigabytes of random-access memory and an Intel Core i5 3.2-GHz processor using $k = 24$, 10,000 queries, and two threads (only for subsampling, the *k*-mer matching portion of Nonpareil is single threaded). Nonpareil alignment kernel results for Iowa soil were obtained by using the Michigan State University High-Performance Computing Center nodes with 20 cores and the following options: 20 threads, 1,000 queries, 50% overlap, and a redundancy to coverage transformation factor of 1.0. All other data sets were processed with a 27-in. iMac as described above and with the same Nonpareil alignment options at 95%

identity and two threads. Iowa soil central processing unit (CPU) time for alignment kernel was estimated by using the linear regression of the other data sets.

N_d. Nonpareil curves are plots of abundance-weighted average coverage ($\hat{C}$) per sequencing effort (*LR*) fitted to the cumulative probability function of the gamma distribution (3) with parameters α and β as follows: $\hat{C} = \gamma[\alpha, \beta \times \log(LR + 1)]/\Gamma(\alpha)$, where Γ is the gamma function and γ is the lower incomplete gamma function. Hence, we can use the mode of the corresponding gamma distribution to identify the value of $\log(LR + 1)$ corresponding to the inflection point of the curve, which we propose as a measurement of sequence diversity as follows: $N_d = (\alpha - 1)/\beta$.

To evaluate the correlation between N_d and traditional measurements of alpha diversity derived from taxonomic affiliation, we compiled three collections of metagenomic data sets. Set I consisted of 90 samples from multiple environments, set II consisted of 54 human-associated microbiome metagenomic samples, and set III consisted of 292 marine samples. Set I was used to evaluate the general agreement between traditional taxonomy-based alpha diversity and N_d. It consisted of 84 metagenomic data sets from multiple environments divided into six distinct biomes plus six mock data sets (Table S3). Subsamples of processed nucleotide reads (between one and seven file chunks of 0.5 gigabyte zipped, as necessary to surpass 60% coverage or as many files as available) and OTU tables based on metagenome-derived 16S rRNA gene-containing reads were obtained from EBI Metagenomics (11). The Shannon index (H') of each OTU table was estimated on the basis of Bayesian estimates of frequencies by using the Dirichlet multinomial pseudocount model with Laplace prior, as implemented in the R package entropy (22), with the exception of mock samples, for which maximum-likelihood entropy of input concentrations was used. Note that the 16S rRNA gene-containing reads derived from metagenomes do not necessarily overlap; hence, the EBI Metagenomics Pipeline uses closed-reference OTU picking (23), potentially biasing the results by database completeness. Therefore, we extended this analysis by using set II derived from the HMP (24). This collection consisted of samples from 13 body sites including 54 metagenomic data sets and 3,613 16S rRNA gene amplicon data sets. Fifty-three samples included both types of data sets. An OTU table including all of the processed samples was obtained from HMP Qiime Community Profiling (23, 24), from which H' values per sample were estimated. Values were compared against N_d values derived from the metagenomic data sets by body site (medians per site) and by sample (intersection samples). Finally, we evaluated the resolution of N_d compared to H' by using set III, a collection of marine samples derived from two global sampling projects, 228 samples from the Tara Oceans expedition between September 2009 and March 2012 (12) and 64 samples from the Global Ocean Sampling expedition between March 2009 and December 2010 funded by the Beyster Family Fund and the Life Technology Foundation (SRA BioProject accession no. PRJEB10418). Sample metadata, as well as preprocessed reads and 16S rRNA gene OTU tables, were obtained from EBI Metagenomics (11). ANOVA was performed to evaluate the effects of different metadata variables on both N_d and Bayesian analysis-corrected H' of extracted 16S rRNA genes. The variables considered were size fraction (categorical), absolute value of latitude (i.e., degrees from the equator), latitude (degrees), geographic location (Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, Indian Ocean, Red Sea, South Atlantic Ocean, South Pacific Ocean, or Southern Sea), sampling date, and two decomposed seasonal components of sampling date. The decomposed seasonal components were estimated as the sine (vernality) and cosine (wintriness) of the date in radians (1 year = 2π) for samples in the Northern Hemisphere or the negative sine and negative cosine, respectively, for samples in the Southern Hemisphere.

Nonpareil 3 availability. Nonpareil 3 is available for online analyses at http://enve-omics.ce.gatech.edu/nonpareil. The Nonpareil code is freely distributed under the artistic license 2.0 and is available at https://github.com/lmrodriguezr/nonpareil.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00039-18.

FIG S1, PDF file, 0.6 MB.

FIG S2, PDF file, 0.3 MB.

FIG S3, PDF file, 0.1 MB.

FIG S4, PDF file, 0.3 MB.

FIG S5, PDF file, 0.5 MB.

FIG S6, PDF file, 0.2 MB.

TABLE S1, PDF file, 0.04 MB.

TABLE S2, PDF file, 0.05 MB.

TABLE S3, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Chengwei Luo and Mark Borodovsky for helpful discussions. We also thank Jiarong Guo for initial test implementations and other testing.

This work has been supported by the U.S. Department of Energy (awards DE-SC0006662 and DE-FC02-07ER64494) and the U.S. National Science Foundation (award 1356288).

REFERENCES

1. Rodriguez-R LM, Konstantinidis KT. 2015. Estimating coverage in metagenomic data sets and why it matters. ISME J 9:1053–1061. https://doi.org/10.1038/ismej.2014.207.
2. Rodriguez-R LM, Castro JC, Kyrpides NC, Cole JR, Tiedje JM, Konstantinidis KT. 2018. How much do rRNA gene surveys underestimate extant bacterial diversity? Appl Environ Microbiol 84:00014–00018. https://doi.org/10.1128/AEM.00014-18.
3. Rodriguez-R LM, Konstantinidis KT. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30:629–635. https://doi.org/10.1093/bioinformatics/btt584.
4. Tamames J, de la Peña S, de Lorenzo V. 2012. COVER: a priori estimation of coverage for metagenomic sequencing. Environ Microbiol Rep 4:335–341. https://doi.org/10.1111/j.1758-2229.2012.00338.x.
5. Větrovský T, Baldrian P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One 8:e57923. https://doi.org/10.1371/journal.pone.0057923.
6. 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. https://doi.org/10.1093/bioinformatics/btp687.
7. Stanhope SA. 2010. Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. PLoS One 5:e11652. https://doi.org/10.1371/journal.pone.0011652.
8. Wendl MC. 2006. A general coverage theory for shotgun DNA sequencing. J Comput Biol 13:1177–1196. https://doi.org/10.1089/cmb.2006.13.1177.
9. Wendl MC, Kota K, Weinstock GM, Mitreva M. 2013. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J Math Biol 67:1141–1161. https://doi.org/10.1007/s00285-012-0586-x.
10. Norling M, Karlsson-Lindsjö OE, Gourlé H, Bongcam-Rudloff E, Hayer J. 2016. MetLab: an in silico experimental design, simulation and analysis tool for viral metagenomics studies. PLoS One 11:e0160334. https://doi.org/10.1371/journal.pone.0160334.
11. Mitchell A, Bucchini F, Cochrane G, Denise H, ten Hoopen P, Fraser M, Pesseat S, Potter S, Scheremetjew M, Sterk P, Finn RD. 2016. EBI metagenomics in 2016—an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res 44:D595–D603. https://doi.org/10.1093/nar/gkv1195.
12. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Ocean plankton. Structure and function of the global ocean microbiome. Science 348:1261359. https://doi.org/10.1126/science.1261359.
13. Luo C, Rodriguez-R LM, Konstantinidis KT. 2013. A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods Enzymol 531:525–547. https://doi.org/10.1016/B978-0-12-407863-5.00023-X.
14. Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239. https://doi.org/10.1016/0888-7543(88)90007-9.
15. Wendl MC, Marra MA, Hillier LW, Chinwalla AT, Wilson RK, Waterston RH. 2001. Theories and applications for sequencing randomly selected clones. Genome Res 11:274–280.
16. Good IJ. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264. https://doi.org/10.1093/biomet/40.3-4.237.
17. Esty WW. 1986. The efficiency of Good's nonparametric coverage estimator. Ann Statist 14:1257–1260. https://doi.org/10.1214/aos/1176350066.
18. Chao A. 1984. Nonparametric estimation of the number of classes in a population. Scand J Stat 11:265–270.
19. Chao A, Shen T-J. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environ Ecol Stat 10:429–443. https://doi.org/10.1023/A:1026096204727.
20. Daley T, Smith AD. 2013. Predicting the molecular complexity of sequencing libraries. Nat Methods 10:325–327. https://doi.org/10.1038/nmeth.2375.
21. Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485. https://doi.org/10.1186/1471-2105-11-485.
22. Hausser J, Strimmer K. 2014. Entropy: estimation of entropy, mutual information and related quantities. https://cran.r-project.org/web/packages/entropy/index.html.
23. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336. https://doi.org/10.1038/nmeth.f.303.
24. Human Microbiome Project Consortium. 2012. A framework for human microbiome research. Nature 486:215–221. https://doi.org/10.1038/nature11209.
25. Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–214. https://doi.org/10.1038/nature11234.
26. Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R, Konstantinidis KT. 2011. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl Environ Microbiol 77:6000–6011. https://doi.org/10.1128/AEM.00107-11.