BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 629–635
doi:10.1093/bioinformatics/btt584

Sequence analysis

Advance Access publication October 11, 2013

Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets

Luis M. Rodriguez-R[1,2] and Konstantinos T. Konstantinidis[1,2,3,*]

[1]Center for Bioinformatics and Computational Genomics, [2]School of Biology and [3]School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Drive, Ford ES&T Building, Suite 3224, Atlanta, GA 30332, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Determining the fraction of the diversity within a microbial community sampled and the amount of sequencing required to cover the total diversity represent challenging issues for metagenomics studies. Owing to these limitations, central ecological questions with respect to the global distribution of microbes and the functional diversity of their communities cannot be robustly assessed.

Results: We introduce Nonpareil, a method to estimate and project coverage in metagenomes. Nonpareil does not rely on high-quality assemblies, operational taxonomic unit calling or comprehensive reference databases; thus, it is broadly applicable to metagenomic studies. Application of Nonpareil on available metagenomic datasets provided estimates on the relative complexity of soil, freshwater and human microbiome communities, and suggested that ~200 Gb of sequencing data are required for 95% abundance-weighted average coverage of the soil communities analyzed.

Availability and implementation: Nonpareil is available at https://github.com/lmrodriguezr/nonpareil/ under the Artistic License 2.0.

Contact: kostas@ce.gatech.edu.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 29, 2013; revised on October 1, 2013; accepted on October 5, 2013

1 INTRODUCTION

Metagenomics have provided important new insights into the diversity, dynamics and functional potential of natural microbial communities during the past decade, but several critical issues remain unresolved. Many metagenomic surveys to date have sampled only a small fraction of the total community DNA; this is particularly the case for soil and sediment communities. Furthermore, the amount of sequencing required to cover the whole community remains speculative (Wendl *et al*., 2012). The fraction of the genomes recovered in a sequencing dataset is termed coverage (Supplementary Box S1), and depends on the sequencing effort applied and the diversity of the community. When the coverage of a metagenome is unknown, results and conclusions about species richness, the evenness of the corresponding community, differences between communities and the extent and importance of rare community members are limited. Moreover, differences between sequencing technologies and continuously changing sequence read lengths make it challenging to establish a universal approach for coverage estimation.

Determining the total number of unique species or operational taxonomic units (OTUs) present in a sample is frequently challenging due to the unknown number of non-sampled species and requires either complete coverage or knowledge of the species abundance curve, which typically remains elusive. The coverage achieved by a dataset can be calculated more efficiently, as it does not depend on *a priori* knowledge of the species abundance curve, and can be directly related to assembly quality (Wendl, 2006). The level of coverage is typically assessed by identifying and counting OTUs and generating rarefaction curves (Hughes *et al*., 2001). Empirical and analytical models have also been applied to coverage estimation using read binning (Hooper *et al*., 2010; Stanhope, 2010) if assembly is not limiting or by targeting specific taxa (Wendl, 2006; Wendl *et al*., 2012) when genome size and abundance are known. However, these approaches and their variations (Hooper *et al*., 2010; Schloss and Handelsman, 2008; Stanhope, 2010; Tamames *et al*., 2012; Wendl, 2006; Wendl *et al*., 2012) require either the use of a reference genome database or the clustering of reads in contigs or OTUs. The former is severely limited by the shortage of representative genome sequences from most habitats (Wu *et al*., 2009). The latter is limited by the quality of the assemblies, especially for highly complex communities, and the use of genes that are much more conserved than the genome average to be sufficiently similar to allow clustering of reads in OTUs such as the ribosomal RNA genes. These genes, however, are known to miss important levels of ecological differentiation among closely related, yet distinct, OTUs (Konstantinidis and Tiedje, 2007). Therefore, a method to estimate the coverage of a metagenomic dataset that is applicable to communities of varied diversity and does not depend on the quality of the assembly and the completeness of reference databases is highly desirable.

Here we introduce Nonpareil ('having no match or equal', referring to the count of unmatched reads in a dataset), a novel method that aims to fulfill this critical gap in contemporary metagenomic research. Nonpareil examines the degree of overlap among individual sequence reads of a whole-genome shotgun (WGS) metagenome to compute the fraction of reads with no match, which is used to estimate the abundance-weighted average coverage (i.e. not the arithmetic mean based on all species in the sample but the average when the abundance of species is considered). Subsequently, it fits a projection line to the estimated values to determine the amount of sequencing required for almost complete diversity coverage. The fraction of unmatched

*To whom correspondence should be addressed.

elements in a given subset of a finite collection (singletons in clustering terms) can be used to efficiently estimate the coverage of the collection, i.e. the fraction of the collection captured in the subset (Esty, 1986; Good, 1953). This observation has been previously applied to metagenomic datasets to estimate species richness (Chao, 1984), functional coverage (Schloss and Handelsman, 2008) and coverage of gene amplicons (Schloss *et al.*, 2009) based on ribosomal RNA or other individual genes. To the best of our knowledge, Nonpareil is the first method directly applying this concept to the whole-genome level, without using reference markers. Further, we propose that Nonpareil projection curves serve as a semi-quantitative proxy to the diversity of the communities. This feature is explored to rank natural communities in terms of the degree of their diversity.

2 METHODS

Our method relies on the observation that datasets with higher coverage are more redundant because the sequencing reads are nearly random, although some systematic biases have been noted for specific sequencing protocols (Dohm *et al.*, 2008). Redundancy is defined here as the portion of reads in a dataset that match with at least one other read (redundant reads; redundant portion is denoted κ). Calculating this value is computationally expensive because it requires a number of paired comparisons asymptotically equal to a quadratic growth (in the worst case, where no two reads match). This is a prohibitive calculation, even for powerful computers, for real-size sequencing datasets that are composed of millions of sequencing reads. Instead, Nonpareil estimates the redundancy value by generating a sample of query reads from the entire dataset (query subset), after which the number of matches per query read in the entire dataset is calculated. For each query read, the total number of matches in the complete dataset is calculated and stored (match-vector; Fig. 1a). Based on the concept of the collector's curve, a saturation function of the redundancy is subsequently produced (Fig. 1b), by iteratively sampling the match-vector in two steps. First, a subset of query reads is selected with a Bernoulli trial per read (with parameter equal to the sampling portion). Next, for each selected query read, the probability of matching another read in the sample is estimated following a binomial distribution, i.e. the number of expected matches of the read in the sample decreases proportionally to the size of the sample, as described in Equation (1).

$$Pr(m \geq 1) = 1 - Pr(m = 0) = 1 - \binom{n}{0} p^0 (1-p)^n$$
$$Pr(m \geq 1) = 1 - \left(1 - \frac{M-1}{N-1}\right)^{(N \times portion)-1} \quad (1)$$

Where m is the number of redundant reads in the subsample, n is the number of reads in the sample, p is the probability of finding a redundant read in the entire dataset, M is the number of redundant reads in the entire dataset, N is the total number of reads in the entire dataset and *portion* is the sampling portion of the entire dataset used for the estimation. This technique prevents redundant comparisons between reads because all comparisons are precomputed once, allowing the calculation of a Nonpareil curve with high resolution (i.e. with sampling portions close to each other). More importantly, it allows multiple replications at each sampling portion (1024 times by default), reducing the effect of randomness in sampling. The resulting function is next summarized (calculating the average, median and standard deviation at each sequencing effort) to estimate the average coverage (Fig. 1c). Finally, a log-gamma regression is fit to the calculated redundancy values with the weighted NL2SOL algorithm (Dennis *et al.*, 1981; Fig. 1d). The projected regression line allows for calculation of the sequencing effort required to reach a fixed average coverage. Nonpareil includes additional optimizations to further decrease the running time and required resources for estimating the redundancy values of datasets of several million of reads on a personal computer, as described later in the text.

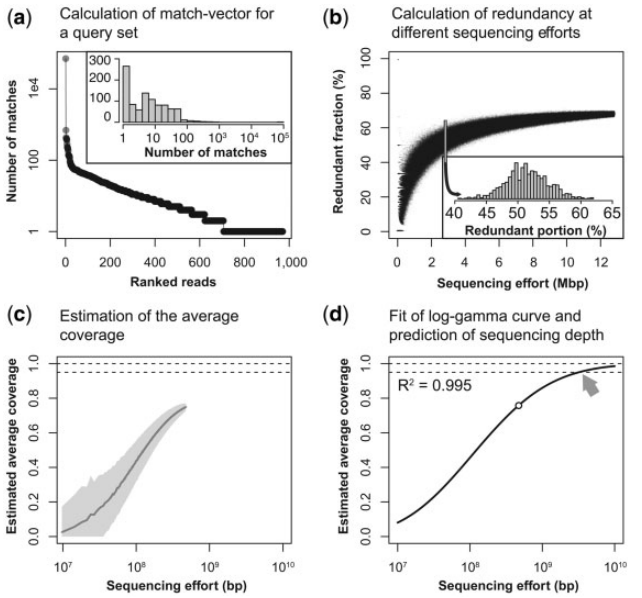

Fig. 1. Main steps in the construction of Nonpareil curves. (**a**) The construction of a Nonpareil curve starts with the calculation of a vector containing the number of matches for a randomly drawn query subset from the total dataset (1000 reads in this case). The function of the number of matches for each query read, ranked by decreasing number of matches, resembles a rank-abundance plot. The inset shows the histogram of matches for the same vector, i.e. the observed distribution of matches from which the rank-abundance plot is generated. (**b**) Next, the redundant portion is calculated for sub-datasets of different sizes. For each size, 1024 replicate datasets are generated. The inset shows the distribution of replicates for a given dataset size. (**c**) The distribution of redundancies is summarized by the average on each size, and the average coverage is estimated. The sequencing effort is displayed in logarithmic scale. The shadowed area represents one standard deviation from the average of the distribution. (**d**) Finally, the curve of estimated coverage is fitted to a log-gamma function and is projected to predict the sequencing effort required to reach a given level of coverage. The solid line represents the fitted function; the empty circle indicates the size of the dataset; and the horizontal dashed lines indicate 100 and 95% coverage. The gray arrow indicates the point where the fitted Nonpareil curve reaches 95% average coverage

2.1 Pairwise read comparison

Nonpareil performs ungapped alignments between reads using a sliding sequence approach that is included in the source code. The alignment strategy aims to match a prefix of the first sequence to a suffix of the second and vice versa. The search space is constrained by the minimum read overlap, excluding comparisons too short to satisfy the threshold, and by minimum identity (default is 95% nucleotide identity; see later in the text), discarding comparisons once the number of allowed mismatches is exceeded. Two sequences are considered redundant (matching pair) if they have at least one alignment satisfying both the minimum overlap and the minimum identity in any strand orientation. Ungapped alignments were preferred because the search space is much smaller than that of gapped alignments, with significant improvements in computation, whereas insertions and deletions between highly identical sequences (i.e. 95% nucleotide

identity or higher) occur at a low frequency. Further, the new sequencers such as the Illumina MiSeq and HiSeq platforms show low rates of insertion and deletion sequencing errors, i.e. <0.01% (Dohm *et al.*, 2008). Finally, sliding sequences are meant to detect overlapping sequencing reads originating from the same genomic region (Wendl, 2006), as opposed to reads with high overall similarity (global alignments) or sharing regions that are not necessarily terminal (local alignments).

2.2 Simulated datasets used in this study

To benchmark our method and resolve the numeric relationship between the redundancy value (κ) calculated by Nonpareil and the average coverage ($\overline{C}$) of a dataset, we generated 120 training datasets by sampling publicly available complete bacterial and archaeal genomes from NCBI's GenBank database uniformly at random (independently of their length and nature). For each dataset, we selected a variable number of genomes, ranging from 1 to 1262. We generated 13 additional datasets using only 282 genomes from *Escherichia coli*, *Yersinia pestis*, *Helicobacter pylori* and *Staphylococcus aureus* to simulate environments with low species richness and phylogenetic diversity, but high intra-species diversity (termed hereafter 'Low richness'). Finally, we produced 10 datasets using 130 genomes from the genus *Escherichia*, simulating environments with extremely low phylogenetic diversity (termed '*Escherichia*'). We randomly assigned an abundance value to each genome in the sample from an exponential distribution ($\lambda = 1$) and produced a number of reads from each genome relative to that value. To produce Illumina-like reads, we randomly selected positions in the genome from a uniform distribution and generated a 101 bp-long read from either strand with randomly introduced sequencing errors from a binomial distribution ($P = 0.01$, $n = 101$). We used the resulting training datasets to evaluate the correlations between Nonpareil indices (κ and R^*) and estimate the average coverage and required sequencing effort for nearly complete coverage (Supplementary Table S1) in log-log space.

2.3 Sequencing depth and coverage estimation

To estimate the coverage of a genome within a training metagenome (generated *in silico*), we backtracked the reads generated from the genome, regardless of the amount of error or the orientation of the reads, and divided the number of covered positions by the genome length. Note that coverage (C) and sequencing depth (ρ; see Supplementary Box S1) share a close relationship, generally approximated through the Lander–Waterman equations (Lander and Waterman, 1988), Equation (2).

$$C = 1 - e^{-\left(\frac{LR\alpha}{\gamma}\right)} = 1 - e^{-\rho} \tag{2}$$

Where LR is the average read length times the number of reads (i.e. the sequencing effort), α is the abundance of the target genome, γ is the length of the genome and ρ is the sequencing depth. To estimate the sequencing depth of a given genome [ρ; Equation (3)], we simply divided the added length of the reads originating from the genome (T) by its length (γ). Note that Equation (2) implies that the number of reads (T) equals the abundance times the number of reads in the dataset (αR).

$$\rho = \frac{TL}{\gamma} \tag{3}$$

Accordingly, we defined average sequencing coverage of a sample as the sum of the sequencing depth of each genome (ρ_i) multiplied by its sequence probability [π_i; Equation (4)].

$$\bar{\gamma} = \sum_i \alpha_i \gamma_i \quad \pi_i = \frac{\alpha_i \gamma_i}{\bar{\gamma}} \quad \bar{C} = \sum_i \pi_i C_i \tag{4}$$

For simplicity, this estimation does not take into account non-observed species because no assumptions about the distribution of abundances of those can be made without additional information. However, species under the detection level are expected to only marginally affect the estimation of both the average coverage and the average genome size because the contribution of each species (i) depends on its abundance (α_i). In cases where the coverage is extremely low, the contribution of non-observed species to the total community can be relatively high, causing unreliable estimates. Nonpareil automatically identifies datasets with estimated coverage below 10^{-5} or with median redundancy of zero in 20% of the subsample and reports them as insufficient data. Although available approximations might provide marginally better estimations of the average sequencing depth (Hooper *et al.*, 2010; Tamames *et al.*, 2012), they rely on assumptions about the shape of the distribution of abundance, which may be unrealistic.

2.4 Estimation of sequencing efforts for nearly complete coverage

To estimate the amount of reads needed to attain a nearly complete coverage of a simulated community/sample, we used Equation 1 to estimate the number of reads necessary to cover 95% of every target genome [R^*_i in Equation (5)] and calculate the average of these values [R^* in Equation (6)]. Note that the value of R for which C equals one is undefined [Equation (2)], and we use $C = 0.95$ as a rule-of-thumb for *nearly complete coverage* [Bouck *et al.*, 1998; *cf.* Wendl *et al.* (2012) for discussion].

$$R^*_i = \frac{-ln(1 - 0.95)\gamma_i}{\alpha_i L} \approx 3\gamma_i/\alpha_i L \tag{5}$$

$$R^* = \sum_i \pi_i R^*_i \tag{6}$$

We define here the sequencing effort required for *nearly complete coverage* of a community (R^*) as the expected number of reads necessary to produce an average coverage of at least 95% of the genomes in all sampled cells.

2.5 Nonpareil curve construction and model fitting

For any given dataset, the Nonpareil curve is defined as the average coverage (estimated from the portion of reads that is similar to at least one other read in the sample; κ) as a function of the sample size ($L{\cdot}R$). Two reads are assumed to be similar if their ungapped alignment shows similarity and length coverage above user-defined thresholds. Here, we used 95% nucleotide sequence identity, intended to reflect natural discrete populations and current species demarcation standards (Caro-Quintero and Konstantinidis, 2012; Goris *et al.*, 2007) and exceed typical sequencing error (Dohm *et al.*, 2008); and three values of alignment length: 25, 50 and 75% of the length of the shortest read. Although we observed that 50% overlap is generally optimal for most datasets, comparisons with 25% overlap should be preferred in extremely low coverage datasets that may be challenging to analyze with 50% overlap. On the other hand, comparisons with 75% overlap may produce fast preliminary results, suitable for high coverage datasets (the longer the overlap, the faster the computation of κ).

The Nonpareil curve has a dual purpose. First, it shows the portion of redundant reads in the entire dataset, reflecting the coverage of the dataset. Second, it allows a projection from the data in hand to the sequencing effort required to achieve any user-defined portion of redundancy, reflecting the complexity of the sample. To perform the projection, the Nonpareil curve is fitted to the cumulative probability function of the gamma distribution [with log-transformed values of the sequencing effort R; Equation (7)].

$$\kappa = \frac{\gamma(a, blog(R+1))}{\Gamma(a)} = \frac{\int_0^{blog(R+1)} e^{-t} t^{a-1} dt}{\int_0^{\infty} e^{-t} t^{a-1} dt} \tag{7}$$

Where Γ is the gamma function, γ is the lower incomplete gamma function (both explicitly noted in the rightmost part of the expression), κ is the redundant portion (coordinate axis in the curve), R is the sample size (ordinate axis in the curve) and a and b are parameters that determine the shape and rate of the curve, respectively, estimated using the weighted NL2SOL algorithm (Dennis *et al.*, 1981).

2.6 Implementation

We implemented the Nonpareil algorithm in C++ with an ancillary R script for model fitting and plotting, using only standard C++ and R libraries. The software parallelizes the read comparisons and sampling steps with an arbitrary number of threads. To reduce the number of hard drive access requests without compromising memory efficiency, blocks of reads are loaded into memory with a maximum random-access memory (RAM) usage defined by the user. This allows the software to run with modest minimal requirements, while ensuring scalability in high-performing computers.

2.7 Real metagenomic datasets

We generated Nonpareil curves for a collection of metagenomic datasets from different environments and levels of diversity. In all cases, we used Nonpareil with default parameters: sequence identity of 95% and read overlap of 50%. We considered the acid mine drainage (AMD) dataset as an example of a community with extremely low phylogenetic diversity (Denef and Banfield, 2012). The sample from site C75 (July 2011), was composed of only *Leptospirillum* sp. group II genotype III, and 5% and 1% subsets of this sample were used to calculate the Nonpareil curves. The genome length of *Leptospirillum* sp. was assumed to be 2.6 megabase (Mb; added length of the scaffolds from GenBank entry AIJM00000000) to calculate both the expected coverage and the expected number of reads required to achieve nearly complete coverage.

We analyzed six selected datasets from the Human Microbiome Project, or HMP (Human Microbiome Project Consortium, 2012), for which both WGS and amplified 16S ribosomal RNA gene (16S) sequencing data were available (Supplementary Table S2). To compare Nonpareil results with a 16S-based estimation, we employed COVER with default parameters (Tamames *et al.*, 2012) to predict the abundance (corrected by 16S copy number) and the genome size of the OTUs in the community from the 16S amplicon data, and used this information to calculate the average sequencing depth and the required effort for nearly complete coverage of the community (Supplementary Table S2). COVER reports the sequencing effort required to achieve a given coverage or a given sequencing depth in the top-n OTUs, but we employed the estimation of R* on Equation (6) (based on abundance and genome length predicted by COVER) to allow comparisons with our method. We also used OTU tables (Caporaso *et al.*, 2010) based on 16S data from http://www.hmpdacc.org/ (Human Microbiome Project Consortium, 2012) to independently assess abundance distributions and Chao1 indexes (Chao, 1984).

We calculated Nonpareil curves for datasets from Lake Lanier (GA, USA), Hess Creek (AK, USA), and the Manu National Park (Peru), representing complex natural environments. We included two samples from August 2009 from Lake Lanier (LL-S1 and LL-S2; Oh *et al.*, 2011), and one additional sample from the same site from July 2010 (LL_1007B) with over twice the sequencing effort, generated with Illumina GA II (100 bp paired-end reads); two soil samples from Hess Creek representing the active and the permafrost layer of the core 2 of day 2 (Mackelprang *et al.*, 2011); and one soil sample from the tropical forest of Manu National Park in Peru (PE6; Fierer *et al.*, 2012). All datasets were trimmed using Solexa QA (Cox *et al.*, 2010) with maximum expected error of 1% and minimum length of 50 bp. In paired-end samples, only the forward reads were used.

3 RESULTS

Nonpareil curves were calculated for 143 short-read simulated metagenomes of various size and diversity levels, generated from publicly available bacterial genomes. Nonpareil estimates of the average coverage of each metagenome correlated strongly (Pearson's $R^2 = 0.93$, $n = 126$) with the independently calculated coverage values based on the known composition of each metagenome. Further, the amount of sequencing that was required to nearly cover the total diversity predicted by Nonpareil (abundance-weighted average coverage of 95% for the genomes in the sample) corresponded tightly to the actual values for each metagenome (Supplementary Fig. S1 and S2 and Supplementary Table S1; Pearson's $R^2 > 0.65$). The estimated abundance-weighted average coverage may also serve as an indicator for the expected quality of the metagenome assembly. Although several factors other than the coverage are critical for assembly, our results show that the average coverage provides a lower-bound estimation of the fraction of assembled reads (Supplementary Fig. S3a), while the assembly N50 of samples with coverage below 60% rarely surpasses twice the read length for Illumina datasets (Supplementary Fig. S3b).

It is important to note that the precision of the algorithm was reduced at values of redundancy (κ) lower than 1% and higher than 90%. These values approximately correspond to <0.01X and >400X sequencing depth, respectively (Supplementary Fig. S4). Because datasets with lower sequencing depth than 0.01X (i.e. too few sequences obtained) are strongly influenced by random variability and thereby subject to spurious results, Nonpareil estimates are not reliable at this range and such datasets are flagged accordingly in the output of the algorithm. Conversely, datasets with sequencing depth above maximum saturation (>400X) are best assessed by read recruitment (mapping), as high-quality assemblies should be achievable in these cases.

3.1 Influence of sequencing error

High frequency of sequencing errors can affect the estimations of the number of redundant reads and thus, Nonpareil curves. It is strongly recommended to filter reads with a stringent cutoff for expected error (e.g. resulting in reads with <1% error rate) prior to applying Nonpareil. The distribution of sequencing error is not always uniform across the length of the reads, depending on the sequencing platform used, and this uneven distribution may affect Nonpareil estimates. In order to evaluate the latter, we analyzed a ~61 Mb dataset generated *in silico* from 33 reference genomes, dominated by *Aeromonas salmonicida* subsp. salmonicida (14%), in which randomly introduced sequencing errors were distributed uniformly, increasing linearly, and increasing as a polynomial of order 4 across the read length (based on Korbel *et al.*, 2009). These involved only wrong-base substitution errors, the dominant source of error in Illumina. The resulting curves (Supplementary Fig. S5) indicated that the estimates of Nonpareil are not affected by the distribution of errors when the total error is ~1% or less, but can be strongly biased when sequencing error approaches 5%. For other types of sequencing errors such as artificial duplicates, a common artifact in 454 sequencing, it is recommended to detect and remove sequences with these errors (e.g. Balzer *et al.*, 2013) prior to applying Nonpareil.

3.2 Coverage estimation of various natural communities

Application of Nonpareil curves to publicly available metagenomes revealed, as expected, that the soil samples required the highest sequencing effort for nearly complete coverage. The seasonally thawed active soil layer from a black-spruce forest in the discontinuous permafrost zone of Alaska at Hess Creek required the highest sequencing effort (Table 1, 0.2 Tb) for nearly complete coverage. The freshwater samples were predicted to require ~10 times less sequencing than soil but more sequencing compared with all evaluated human microbiota (e.g. >80X more than posterior fornix; Fig. 2 and Table 1). The AMD sequences mapping to *Leptospirillum* sp. covered 99.99% of the genome, and the 1% subset covered 73%. Using Nonpareil, a sequencing coverage of 70% was estimated for the 1% subset, which corresponds to 17 Mb of the complete dataset, whereas an expected coverage of 94% was obtained using the Lander–Waterman expression [Lander and Waterman, 1988; Equation (2)].

In addition to WGS metagenomes, the HMP samples included amplified 16S ribosomal RNA gene (16S) sequencing data (Supplementary Table S2). Estimation of the abundance and genome size of 16S-defined OTUs by COVER (Tamames *et al*., 2012) displayed larger variability (*cf*. Table 1 and Supplementary Table S2) compared with Nonpareil estimates based on metagenomes from the same samples, possibly reflecting the influence of sequencing errors or polymerase chain reaction artifacts (Kunin *et al*., 2010) or variations on the assumed number of 16S copies (Větrovský and Baldrian 2013). The largest difference between COVER and Nonpareil was observed in the posterior fornix, an environment known to be largely dominated by *Lactobacillus* (Ravel *et al*., 2011). Quality-checked 16S data showed that the most abundant OTU accounted for ~90% of the community and only nine of 37 OTUs identified showed abundance >0.1%. Assuming a typical vaginal lactobacilli genome size of 2.4 Mb, ~8 Mb are predicted by the Lander–Waterman expressions (Lander and Waterman, 1988) to be required to cover 95% of the dominant OTU. Nonpareil estimated that 12 Mb of sequence data provide an average coverage of 91%, and ~40 Mb would be necessary to cover the community almost entirely. In contrast, COVER estimated a total of 226 OTUs and 160 Gb to be necessary for the same level of coverage. These results suggest that 16S analysis can frequently inflate diversity, resulting in large underestimations of the sequencing depth. They also reveal that Nonpareil produces estimations closer to those of other genome-wide approximations such as the Lander–Waterman model. We limited our evaluation to COVER because alternative methods for coverage estimation (Daley and Smith, 2013; Hooper *et al*., 2010; Stanhope, 2010; Wendl *et al*., 2012) were either not available for online or standalone computation, do not scale with large metagenomic

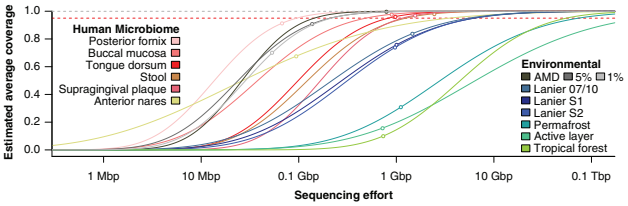

Fig. 2. Comparison of Nonpareil curves for the metagenomes of HMP, AMD, Lake Lanier, Permafrost soil and Tropical Forest soil. The plot displays the fitted models of the Nonpareil curves. The horizontal dashed lines indicate 100 (gray) and 95% (red) coverage. The empty circles indicate the size and estimated average coverage of the datasets, and the lines after that point are projections of the fitted model. The curves cluster in four groups, reflecting different levels of diversity. The leftmost group, including posterior fornix, buccal mucosa, anterior nares and AMD, represents samples largely dominated by a single species. A second group, composed by tongue dorsum, stool and supragingival plaque, represents low to medium diversity samples. Next, freshwater samples, which are typically characterized by moderate to high diversity, cluster together. Finally, the curves for the high-diversity soil samples display the lowest slopes

Table 1. Nonpareil estimates for publicly available metagenomic datasets

Location	Identifier	$L{\cdot}R$	$\overline{C}$ (%)	L·R*	Data source
Anterior nares	SRS019087	20 Mb	68	2.3 Gb	(Human Microbiome Project Consortium, 2012)
Buccal mucosa	SRS063287	1.0 Gb	95	–	
Stool	SRS016335	5.6 Gb	97	–	
Supragingival	SRS015574	2.5 Gb	97	–	
Tongue dorsum	SRS062540	1.2 Gb	95	–	
Posterior fornix	SRS063417	12 Mb	91	43 Mb	
Richmond mine (CA, USA)	C751107	0.7 Gb	99	–	(Denef and Banfield, 2012)
	C751107 1%	7.2 Mb	70	90 Mb	
Lake Lanier (GA, USA)	LL-S1	1.1 Gb	72	3.2 Gb	(Oh *et al*., 2011)
	LL-S2	1.1 Gb	73	3.1 Gb	
	LL_1007B	2.3 Gb	83	2.8 Gb	This study
Hess Creek (AK, USA)	Permafrost C2	1.8 Gb	31	41 Gb	(Mackelprang *et al*., 2011)
	Active C2	1.6 Gb	16	198 Gb	
Manu Park (Peru)	PE6	0.7 Gb	10	30 Gb	(Fierer *et al*., 2012)

Note: All analyses were executed with default parameters (50% read overlap, 95% identity). For each sample, Nonpareil estimated the average coverage ($\overline{C}$) and predicted the sequencing effort required for nearly complete coverage ($L{\cdot}R^{*}$). Dashes (-) indicate that the model was not projected because the estimated coverage exceeds 95%. Identifiers starting with SRS indicate entries in the NCBI Sequence Read Archive; all other identifiers are from to the original publications.

datasets or provide results like longest contig expected per taxon that are not directly comparable with those of Nonpareil.

3.3 Diversity ranking

An interesting feature of the Nonpareil curves is that the shape of the curves reflects the level of diversity of the communities sampled. The Nonpareil curve saturates faster, i.e. complete coverage is achieved with fewer sequences sampled, on datasets with lower diversity and shorter genomes (Fig. 2). Because the average genome sizes differ by no more than one order of magnitude between most microbial communities, the velocity of saturation of Nonpareil curves is mostly determined by the sample diversity rather than differences in genome size or gene duplications and repetitive regions. However, deviations from this expectation are possible when comparing metagenomes with large differences in average genome size, as it is often the case when the proportions of viral, bacterial/archaeal and eukaryotic DNA differ substantially. In such cases, separation of the different fractions (e.g. Liu *et al*., 2013) before applying Nonpareil is recommended and the efficiency of this technique needs to be assessed on a case by case basis. Figure 2 revealed clustering of curves from samples with decreasing levels of diversity. Nonpareil curves from samples of communities characterized by low diversity, like posterior fornix, anterior nares and AMD, rapidly saturated. In contrast, Nonpareil curves from soil samples, known to possess comparatively high diversity, continued growing after projecting to millions of reads. Intermediate in Figure 2 are Nonpareil curves from freshwater samples, stool and tongue dorsum, expected to have a higher diversity than the first group of samples but lower than soil datasets. This property of the Nonpareil curves allows fast assessment of the level of diversity inherent to an unknown sample compared with reference communities. In addition, the shape of the Nonpareil curves can reveal distinctive features of the samples such as skewed distribution of species abundances. For example, the Nonpareil curve for the anterior nares sample (Fig. 2) showed a rapid growth phase at low sequencing effort that does not saturate as rapidly as other low-complexity samples. Further examination indicated that this pattern was due to an unusual distribution of abundances (as revealed by 16S profiling; Human Microbiome Project Consortium, 2012), following an extreme broken-stick model. In all, 74 species were observed and ~99 species were estimated to coexist in this sample (Chao1, $IC_{95\%}$: 32.98–239.7) but the most abundant species had an abundance of 36%, and the 9 most abundant species represented 95% of the community.

3.4 Computing performance

We tested Nonpareil with datasets of various sizes (101 bp-long reads) and evaluated its performance in terms of central processing unit (CPU) time, running time and RAM usage (Supplementary Fig. S6). All tests were performed on cluster architecture with 64 CPUs (2.2 GHz) per node, >40 GB of available RAM, running on Red Hat Enterprise Linux 6. Both the running time and the RAM usage grow linearly with the size of the dataset, as anticipated. The RAM use in GB was ~0.1 times the size of the dataset (in millions of reads) plus 2. This relationship might vary on different computers, operating systems and future versions of the code, but it offers an indication of the RAM requirement of the algorithm without parceling. Note that the maximum RAM usage can be set on each run by the user, and Nonpareil can parcel the data to adapt to less powerful computers as needed. Both the running time and the CPU time are strongly affected by the stringency (cutoffs) of the read comparison (Supplementary Fig. S6b and c). However, the algorithm scaled up equally well with all the parameters (Supplementary Fig. S6d).

4 CONCLUSIONS

The results presented here highlight the usefulness of the Nonpareil curve as a tool for both study design and exploratory comparisons of community diversity. This tool increases the range of samples for which coverage can be computed relative to existing tools. It is important to point out that existing approaches for coverage estimation require prior knowledge about the abundance distribution of the members of the community (Wendl, 2006; Wendl *et al*., 2012) and/or assume that the diversity distribution can be effectively modeled by known probability distributions (Hooper *et al*., 2010; Stanhope, 2010), or require the use of reference molecular markers (Daley and Smith, 2013; Tamames *et al*., 2012). These properties of a metagenome are frequently not available. The relationship between sequencing effort and average coverage of the community can be alternatively approximated by visual inspection of rarefaction when binning is feasible (Schloss and Handelsman, 2008; Schloss *et al*., 2009). A recent development improved on this traditional approach by providing a mathematical generalization for any molecular marker and an accurate projection of the rarefaction curve (Daley and Smith, 2013). However, the level of coverage remains inaccessible and sequence binning is a required step, which is typically limiting in WGS metagenomic studies. In contrast, Nonpareil does not require abundance distributions, models or reference databases and is based on the redundancy of the reads, an intrinsic characteristic of any metagenomic dataset. The complement of redundancy is the number of reads without matches in a given sample divided by the sample size, which we denoted as the Nonpareil fraction (υ). When expressed in terms of non-matching reads (i.e. one minus the Nonpareil fraction) the redundancy essentially takes the same form as the Good's coverage estimator (Good, 1953), a widely applied estimator of coverage of a sample (Esty, 1986). Nonpareil applies this estimation directly on shotgun sequencing reads, even in datasets composed of millions of reads, with modest computational requirements.

Application of Nonpareil estimates on available metagenomes revealed, as expected, that the largest sequence efforts were required for soil datasets, where up to 200 Gb and 1 Tb of sequence data were predicted to be necessary to achieve 95 and 99% abundance-weighted average coverage, respectively. These estimates are well below the 10 Tb estimate of Riesenfeld *et al*. (2004) required to cover a typical soil metagenome, which emphasizes on coverage of all species, including rare ones. For example, Nonpareil predicts an increase in average coverage from 99.9 to 99.99% with 1–10 Tb of data (in Hess Creek), a marginal difference in abundance-weighted average coverage for 10 times more data. These results agree with previous findings based on single target species (Wendl *et al*., 2012), supporting that the estimations of Nonpareil are practical and robust. The soil dataset of Hess Creek represents a permafrost soil incubated under warm

temperatures, which likely stimulated specific taxa, affecting the diversity of the community. However, the Manu Park sample represents a temperate soil, estimated to contain close to 9000 species based on 16S data (from 5347 observed 97% OTUs; Fierer *et al*., 2012). In fact, the estimate provided by Nonpareil for 99% average coverage (95 Gb) translates to complete coverage of any genome of ~5 Mb with abundance >0.07 with 90% confidence (Wendl *et al*., 2012). This corresponds to the top 312 most abundant species, or 90% of the observed community, based on 16S (Fierer *et al*., 2012). Note that these 312 species likely represent only ~5% of the number of species present in the community. However, Nonpareil estimate is not meant to reflect the captured richness of the community (i.e. how many different species were captured), but the portion of the total community captured, taking abundance into consideration.

Finally, we evaluated the robustness of Nonpareil estimates by both decreasing the sequencing effort on a community with high coverage (AMD) and increasing it on a community with medium coverage (Lake Lanier). In both cases the estimates were consistent with the expectations (Fig. 2 and Table 1), indicating that Nonpareil analysis is robust to variations in the size of the query dataset, and variations arising from independently collected samples or different sequencing protocols (Lake Lanier samples).

In summary, Nonpareil curves offer an estimation of average coverage of metagenomic datasets (for profiling studies and other community-wide analyses), a prediction of coverage in increased sequencing efforts (for study design) and a comparative framework for diversity exploration, allowing for fast diversity rankings of metagenomes before assembly or taxonomic classification.

5 AVAILABILITY

Nonpareil is free software licensed under the terms of the Artistic license 2.0. The source code and binaries are available at https://github.com/lmrodriguezr/nonpareil/. An online server is available at http://enve-omics.ce.gatech.edu/nonpareil/. Sequences of Lake Lanier (LL_1007B) were deposited in the NCBI Sequence Read Archive, with accession number SRR948155.

ACKNOWLEDGEMENTS

The authors thank Heidi Kizer, Janet Hatt and three anonymous reviewers for helpful suggestions regarding the manuscript.

Funding: U.S. Department of Energy (Award DE-SC0006662) and by U. S. National Science Foundation (Award No 1241046).

Conflict of Interest: none declared.

REFERENCES

Balzer,S. *et al*. (2013) Filtering duplicate reads from 454 pyrosequencing data. *Bioinformatics*, **29**, 830–836.

Bouck,J. *et al*. (1998) Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res*., **8**, 1074–1084.

Caporaso,J.G. *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Caro-Quintero,A. and Konstantinidis,K.T. (2012) Bacterial species may exist, metagenomics reveal. *Environ. Microbiol*., **14**, 347–355.

Chao,A. (1984) Nonparametric estimation of the number of classes in a population. *Scand. J. Stat*., **11**, 265–270.

Cox,M.P. *et al*. (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, **11**, 485.

Daley,T. and Smith,A.D. (2013) Predicting the molecular complexity of sequencing libraries. *Nat. Methods*, **10**, 325–327.

Denef,V.J. and Banfield,J.F. (2012) *In situ* evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science*, **336**, 462–466.

Dennis,J.E. *et al*. (1981) An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Softw*., **7**, 348–368.

Dohm,J.C. *et al*. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res*., **36**, e105.

Esty,W.W. (1986) The efficiency of good's nonparametric coverage estimator. *Ann. Stat*., **14**, 1257–1260.

Fierer,N. *et al*. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl Acad. Sci. USA*, **109**, 21390–21395.

Good,I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Goris,J. *et al*. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol*., **57**, 81–91.

Hooper,S.D. *et al*. (2010) Estimating DNA coverage and abundance in metagenomes using a gamma approximation. *Bioinformatics*, **26**, 295–301.

Hughes,J.B. *et al*. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol*., **67**, 4399–4406.

Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Konstantinidis,K.T. and Tiedje,J.M. (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr. Opin. Microbiol*., **10**, 504–509.

Korbel,J.O. *et al*. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*., **10**, R23.

Kunin,V. *et al*. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol*., **12**, 118–123.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Liu,J. *et al*. (2013) Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res*., **41**, e3.

Mackelprang,R. *et al*. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, **480**, 368–371.

Oh,S. *et al*. (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.

Ravel,J. *et al*. (2011) Vaginal microbiome of reproductive-age women. *Proc. Natl Acad. Sci. USA*, **108** (**Suppl. 1**), 4680–4687.

Riesenfeld,C.S. *et al*. (2004) Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet*., **38**, 525–552.

Schloss,P.D. and Handelsman,J. (2008) A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, **9**, 34.

Schloss,P.D. *et al*. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol*., **75**, 7537–7541.

Stanhope,S.A. (2010) Occupancy modeling, maximum contig size probabilities and designing metagenomics experiments. *PLoS One*, **5**, e11652.

Tamames,J. *et al*. (2012) COVER: a priori estimation of coverage for metagenomic sequencing. *Environ. Microbiol. Rep*., **4**, 335–341.

Větrovský,T. and Baldrian,P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One*, **8**, e57923.

Wendl,M.C. (2006) A general coverage theory for shotgun DNA sequencing. *J. Comput. Biol*., **13**, 1177–1196.

Wendl,M.C. *et al*. (2012) Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. *J. Math. Biol*., **67**, 1141–1161.

Wu,D. *et al*. (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and *Archaea*. *Nature*, **462**, 1056–1060.