ISME Communications, 2025, 5(1), ycaf140
https://doi.org/10.1093/ismeco/ycaf140
Advance access publication: 10 September 2025
Original Article

Differences in metagenome coverage may confound abundance-based and diversity conclusions and how to deal with them

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

[1]School of Civil & Environmental Engineering and School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, United States
[2]Department of Biomedicine and Dentistry, European University of Andalucia, Málaga, 29010, Spain
[3]Department of Microbiology & Digital Science Center (DiSC), University of Innsbruck, Innsbruck, A-6020, Austria

*Corresponding authors. Borja Aldeguer-Riquelme, Department of Biomedicine and Dentistry, European University of Andalucia, Eolo Street 12, Malaga, 29010, Spain. E-mail: briquelme3@gatech.edu; Luis M. Rodriguez-R, Department of Microbiology & Digital Science Center (DiSC), University of Innsbruck, Innrain 15, A-6020 Innsbruck, Austria. E-mail: lmrodriguezr@gmail.com; Konstantinos T. Konstantinidis, Ford Environmental Sciences and Technology, School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr NW, Atlanta, GA 30332, USA. E-mail: kostas@ce.gatech.edu.

Abstract

The importance of rarefying ecological or amplicon sequencing data to a standardized level of diversity coverage for reliable diversity comparisons across samples is well recognized. However, the importance of diversity coverage, i.e. the fraction of the genomic diversity of a sample sequenced, in comparative shotgun metagenomic studies remains frequently overlooked. Using both *in silico* and natural metagenomes from a wide range of environments, we demonstrate that uneven metagenome coverage can result in misleading biological conclusions, particularly for identifying differentially abundant features, i.e. groups of genes or genomes assigned to the same protein family or taxonomic rank, respectively, and for comparing diversity between samples. The main underlying cause is that not all members of a feature may be detectable, and thus counted, across such unevenly covered metagenomes depending on the sequencing effort applied and the underlying member-abundance curves. Unfortunately, 99.5% of previous comparative metagenomic studies have overlooked this metric, suggesting that their reported results might be misleading. We show that achieving high Nonpareil coverage (≥0.9), a metric that estimates metagenome diversity coverage, is the most reliable strategy to mitigate this issue. When high Nonpareil coverage is not achievable, such as for highly diverse and complex samples like soils, we show that standardizing (or subsampling) metagenomic datasets to the same Nonpareil coverage, rather than sequencing effort, prior to comparative analysis provides for more accurate results. We provide a set of practical recommendations and the corresponding Python scripts to help researchers to assess and standardize metagenome diversity coverage for their comparative analyses.

Keywords: metagenome; coverage; differential abundance; diversity; nonpareil

Introduction

Modern microbiome studies frequently employ sequencing approaches, such as amplicon sequencing or shotgun metagenomics, to answer ecological-, evolutionary-, or diversity-related questions. However, only a fraction of the whole microbial community is typically sequenced in these studies. Therefore, determining the amount of diversity recovered by the sequencing effort applied (hereinafter, diversity coverage) is essential for understanding the representativeness of the dataset and the reliability of the conclusions obtained from it. That is, significant deviations in the diversity coverage between datasets might distort the inferred differences between the corresponding communities in terms of their phylogenetic or functional diversity [1]. Strategies to estimate the diversity coverage of a dataset usually rely on a redundancy metric or the detection of single occurrences (singletons) of diversity units, such as Operational Taxonomic Units (OTUs) or reads. These methods provide both qualitative (e.g. rarefaction curves) [2] and quantitative estimates (e.g. Turing–Good's coverage) [3] of diversity coverage.

Traditional rarefying analysis attempts to account for (or standardize) uneven diversity coverage between samples by subsampling all datasets to the same number of sequences or/and diversity of features (e.g. % of total OTUs). Rarefying has been shown to be necessary when comparing 16S rRNA gene amplicon datasets of varied diversity coverage or number of sequences in order to infer differentially abundant taxa, differences in diversity and more between the datasets [4–6]. This standardization strategy has been widely adopted in the literature due to its implementation in popular tools for analyzing 16S rRNA gene amplicon data, such as Mothur or QIIME2 [7, 8]. Moreover, it is now well recognized that rarefying amplicons by sequencing effort (i.e. number of sequences used) is an insufficient, and often counterproductive method [1, 9], and diversity coverage standardization should be instead used [10–12]. There are several reasons

Received: 25 June 2025. **Revised:** 17 July 2025. **Accepted:** 8 August 2025
© The Author(s) 2025. Published by Oxford University Press on behalf of the International Society for Microbial Ecology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

supporting this claim. First, samples standardized by size may have different degrees of diversity coverage and thus, representativeness, depending on their species richness and species-abundance distributions. In a simplified example, a sample with 1,000 species will require around 10X more sequencing effort to be as well covered as a sample with only 100 species and the same degree of evenness. Second, standardizing samples by size does not satisfy the replication principle while diversity coverage-based standardization does [10, 11]. The replication principle ensures that the ratio of diversity richness between samples is maintained when comparing their subsampled datasets [10]. Thus, by preserving this principle, the diversity coverage-based standardization approach better reflects the real differences in diversity between samples or their subsamples than the size-based standardization. Third, diversity coverage-based projections are finite because this variable is bound between 0 and 1, unlike size-based projections. This is important because it allows computing asymptotic estimates of diversity and (some) closed-form confidence intervals [13]. Despite the recognized importance of diversity coverage in ecological and amplicon studies, this metric is still largely ignored in comparative shotgun metagenomic studies. This is presumably because it is technically much more challenging to estimate diversity coverage with (largely) non-overlapping sequences like in shotgun metagenomic datasets (or simply metagenomes) relative to overlapping amplicon data.

Rodriguez-R and Konstantinidis developed Nonpareil in 2014 to calculate the diversity coverage of a metagenome based on read redundancy [14] (Fig. 1). Nonpareil coverage (Npc) is calculated by measuring the fraction of redundant reads in subsets of different sizes generating a redundancy curve, which is then fitted to a log-gamma distribution to estimate the sequencing effort needed to reach a given level of metagenome coverage [14]. Thus, by measuring the metagenome diversity coverage, Npc can be considered a proxy of sample diversity coverage. Npc does not correlate linearly with sequencing effort but instead depends on the biological complexity of the community, including factors such as species evenness and richness [14]. As a result, Npc represents a biologically-informed metric, in contrast to sequencing effort, which is a purely technical metric. Npc is commonly referred in the metagenomic literature as "coverage", which can generate some confusion with other metrics that employ the same term, such as sequencing depth (SD) and breadth, also referred to as coverage depth and breadth, respectively. However, these are three distinct metrics that should not be conflated. Npc reflects the proportion of diversity captured in a metagenomic dataset, whereas SD and coverage breadth are metrics derived from analysis of read mapping against a reference sequence (Fig. 1). SD quantifies the frequency of reads mapped on the reference (e.g. 5X depth, or five sequencing reads per position on average), while coverage breadth measures the extent to which a reference sequence is covered by mapped reads (e.g. 50% breadth, or half of the length of the reference covered by reads; Fig. 1). To prevent confusion, we do not use the term "coverage" alone, but in conjunction with the appropriate type; i.e. diversity coverage, coverage breadth, and coverage depth. Additionally, to clarify potentially confusing concepts used throughout this manuscript, we define their meanings in Table 1.

The Nonpareil algorithm was significantly improved in 2018. The advancements included the use of k-mers to reduce the runtime of large metagenomes, the implementation of a sequencing error correction, and the introduction of the Nonpareil diversity metric [15]. A preliminary study was also conducted to explore the impact of Npc on gene detection and differential abundance analyses between metagenomes and concluded that two metagenomes should not be compared if they show more than 2-fold difference in metagenome coverage, otherwise spurious results could emerge [16]. However, the study included only two real, incomplete metagenomes with Npc of 0.64 and 0.75; a more comprehensive evaluation under different scenarios of Npc differences among datasets remained elusive. Further, the absence of an available ground-truth reference (i.e. the real abundances of features in the sequenced library to be known prior to the analysis) and the limited number of samples used significantly constrained the generalizability of the findings. Accordingly, while diversity coverage has been proposed as a standardizing strategy for ecological studies [10], exploring the potential of Npc to standardize metagenomic datasets warrants further investigation.

Standardization approaches of metagenomes have been largely limited to correct for uneven SD, either using methods borrowed from transcriptomics or rarefying [17]. However, the transcriptomics methods typically have assumptions that are violated by common metagenomes (see Supplementary Note 1). Rarefying, despite being likely the most common approach, showed a low performance for identifying differential abundance genes between metagenomes [17]. Therefore, more robust standardization methods specifically designed for metagenomics are needed to accurately normalize relative abundance values.

Here, we systematically evaluated the impact of uneven diversity coverage on metagenomic comparisons using both *in silico* and real metagenomes from natural environments. We further highlight the importance of metagenome coverage by showing that overlooking coverage differences has led several published studies to inaccurate conclusions about what features (e.g. genes, pathways, or taxa) differ in abundance and/or diversity (e.g. number of species) between microbiomes. Note that the recovery of microdiversity (e.g. the number of single nucleotide polymorphisms [SNPs] within a species) as a function of the sequencing effort (depth) applied is a distinct issue (than the macrodiversity we focus herein), and is dealt with elsewhere [18]. While previous Nonpareil studies primarily focused on software development, here we expand its use cases by introducing Npc coverage standardization as an approach to mitigate the effects of differences in diversity coverage between datasets on comparing feature abundances or diversities. We provide a straightforward decision tree to guide researchers in conducting robust feature abundance comparisons between metagenomic datasets, and explain why standardizing to the same sequencing effort (number of reads) is frequently not as robust as Npc standardization.

Materials and methods

Construction of in silico metagenomes

In silico (synthetic) metagenomes with known composition (ground truth) were simulated controlling for different biological characteristics, including evenness (even vs uneven), species richness (100 vs 1000 species) and microdiversity (1 vs 10 genome per species). Genomes were downloaded from NCBI and GTDB databases (Table S1). Short read sequences (150 bp) were simulated from genomes using Mason v2.0.9 [19] (–fragment-mean-size 150 –seq-technology illumina), which is known to accurately replicate the features of modern Illumina sequencers [20]. The number of reads per genome was calculated and the reads were subsequently subsampled appropriately to ensure that the resulting taxon rank curves conformed to a log-normal distribution, reflecting the patterns typically observed in real

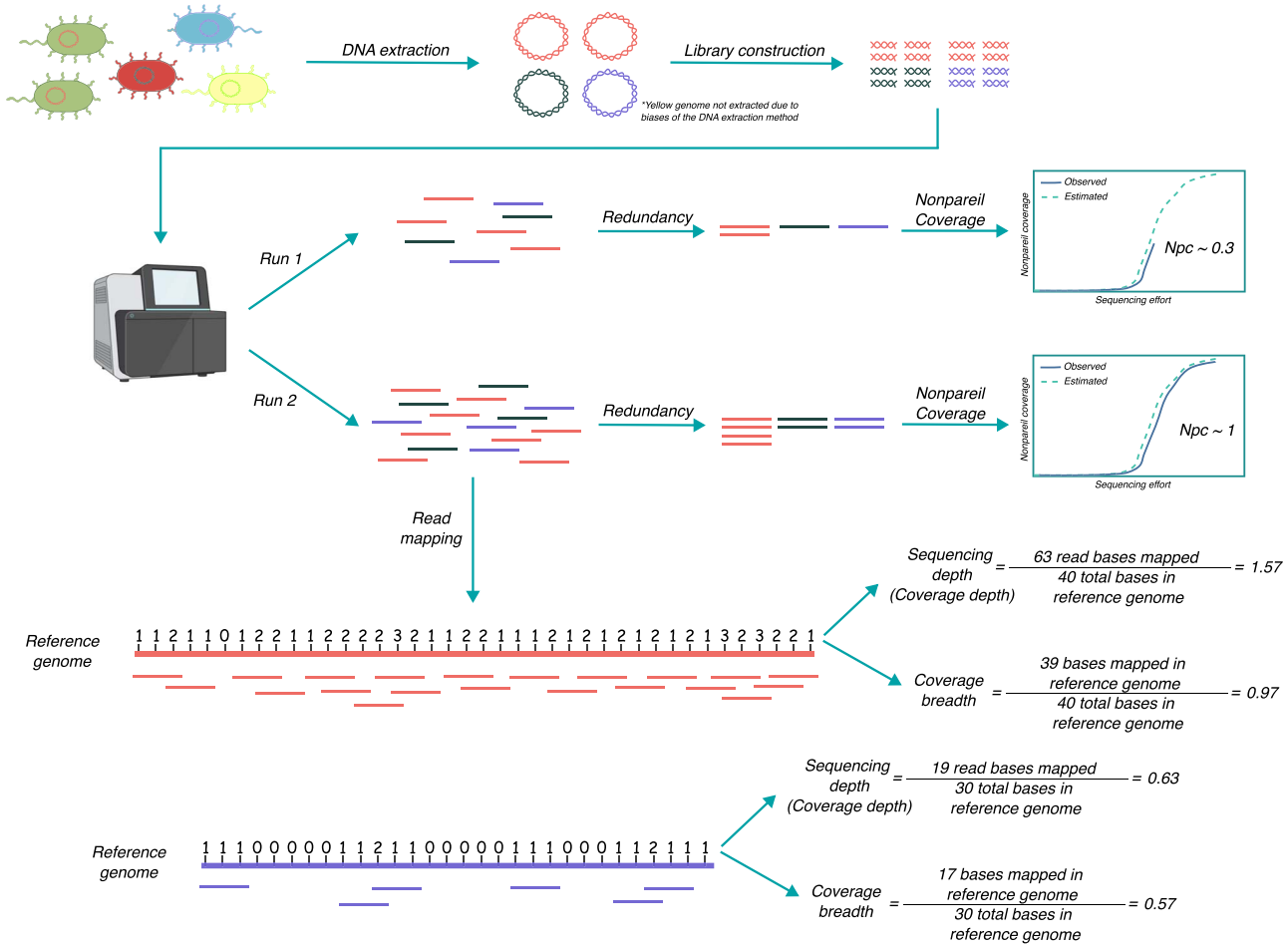


Figure 1. Schematic representation of the Npc calculation and its differences from sequencing depth (SD; or coverage depth) and sequencing breadth (or coverage breadth). All three metrics are commonly referred to as "coverage" in the literature, causing confusion about their exact meaning or differences. As depicted in the figure, Nonpareil is a database-independent approach as opposed to SD and coverage breadth that measure the frequency and extent of the reads mapped to a reference sequence, respectively. Note that Npc is calculated based on the redundancy of sequenced reads, with higher Npc values as redundancy increases. DNA present in the sample but not made into the library for sequencing (represented by the yellow genome) will not be accounted for in the Npc calculation. Partially created in BioRender. Aldeguer, B. (2025) https://BioRender.com/qmrhjlb.

Table 1. Definitions of potentially confusing terms employed in this manuscript.

Term	Definition
Diversity coverage	Fraction of the total diversity recovered in a sample, typically expressed as the fraction of cells belonging to species detected in the sample [range: 0–1]
Metagenome coverage	Fraction of the total DNA diversity recovered in a metagenome (proxy for diversity coverage) [range: 0–1]
Sequencing depth (or coverage depth)	Average number of times a given reference sequence is mapped by metagenomic reads [range: 0 – Inf]
Sequencing breadth (or coverage breadth)	Fraction of the reference sequence mapped by metagenomic reads [range: 0–1]
Rarefying	Process of subsampling a dataset once to obtain a single measurement
Rarefaction	Process of subsampling a dataset multiples time to obtain a central measurement
Normalization	Process of transforming a set of values relative to a given factor to ensure consistency, comparability and/or scale adjustment (e.g. divide sequencing depth by genome equivalents to calculate relative abundance).
Standardization	Process of estimating a given metric using consistent methods under uniform conditions to ensure comparability across datasets (e.g. estimate relative abundance at the same Npc).

metagenomes [21]. All reads subsampled for each genome were then merged into a single file representing the *in silico* metagenome. The evenness of the distribution was determined by the location parameter (μ) of the log-normal equation (two and five for uneven and even distributions, respectively) and the ratio between the maximum and minimum value. Ratios of 10,000 and 500 were applied to get uneven and even distributions, respectively, based on the values we previously observed in real metagenomes. Note that for low abundance genomes the number of reads may be so low that only a fraction of the genome is represented in the final metagenome, similarly to real metagenomes. In order to facilitate reproducibility, the pipeline described here was

implemented in the script "MetaG_simulator.py", which is publicly available at https://github.com/baldeguer-riquelme/Nonpareil-coverage-standardization.

Genome detection and relative abundance estimation

Detection of genomes/metagenome-assembled genomes (MAGs)/genes was based on read mapping analysis (i.e. assembly-independent). Specifically, relative abundance was calculated as SD divided by Genome Equivalents (GEQ), which normalizes for differences in the sequencing effort applied, length of MAGs/targets, and average genome size between the metagenomes [22] as well as for spurious matches [23, 24]. For this purpose, SD and GEQ were estimated by CoverM v0.6.1 (genome -p bwa-mem –min-read-percent-identity 95 –min-read-aligned-percent 70 –min-covered-fraction 10 –exclude-supplementary -m mean) [25] and MicrobeCensus v1.1.0 (default settings) [22], respectively. CoverM was also used to calculate mapped read counts which were converted to relative abundance as Read Per Kilobase per Million reads (RPKM). For simplicity, we used "relative abundance" to refer to the SD/GEQ metric, unless otherwise noted; e.g. for RPKM data. Plots were drawn in R using the ggplot2 library [26].

Nonpareil coverage standardization

Metagenomes were subsampled to Npc values ranging from 0.1 to 0.9 in 0.1 increments. For this, the number of reads needed to reach a given Npc was calculated with the predict.Nonpareil.Curve function from the Nonpareil R library v3.5.3 using the npo files from Nonpareil v3.4.1 [15]. Then, reads were randomly subsampled to the target number of reads using reformat.sh. The approach was implemented in the Npc_standardization_manual.R script. Users can use this script to get randomly subsampled metagenomes with a given Npc and subsequently perform read mapping to features, etc. Alternatively, to simplify and speed up Npc standardization and reduce estimation error, relative abundance or count values can also be estimated using the Npc_standardization.R script. The .npo files from the Nonpareil analysis, the original relative abundance, GEQ or count values and the MAG/gene length should be provided to run the script. The proposed standardization approach consists of the following steps:

1. Estimation of the fraction of reads required to achieve a given Npc for each metagenome. For this purpose, the predict.Nonpareil.Curve function is used to calculate the number of reads at the target Npc, which is then divided by the total number of reads in the original metagenome.

2. The fraction obtained is multiplied by SD, GEQ or reads counts in the original metagenome to get the estimated relative abundance values in the subsampled metagenomes.

3. Finally, only MAGs/genes with a minimum 0.1X SD, which equates to ∼10% coverage breadth according to the Lander–Waterman equation [27], are considered as present and used for comparisons between metagenomes. The remaining MAGs/genes are considered as undetected (i.e. zeros are assigned to MAGs/-genes with SD <0.1X).

Richness and differential abundance tests

Richness was assessed with three metrics, namely number of detected species (i.e. SD/GEQ > 0), Shannon index and Simpson index. Both Shannon and Simpson indexes were obtained with the diversity function of the vegan R package v2.6.8 [28]. Regarding differential abundance tests, the aggregated relative abundance of genomes belonging to the same phylum was compared between two metagenome types using the t.test() function from the stats R package. The significance of the statistical test as well as the direction (i.e. higher abundance in metagenome A than metagenome B or vice versa) in the subsampled metagenomes was compared to that obtained with the original metagenomes. When both results matched, the differential abundance test was labeled as accurate.

Maximum acceptable difference in Npc (ΔNpc_{max})

To calculate the ΔNpc_{max} in a systematic fashion, we used the t.test function from stats R package v4.2.0 to compare the relative abundance at 0.7 Npc (reference) of each taxon at the phylum, class, order, and family level against the relative abundance of the same taxon in subsampled metagenomes showing stepwise decreasing Npc values (0.01 Npc per step). We ran this loop until the t.test became statistically significant. The difference between 0.7 and the specific Npc of the subsample considered was termed ΔNpc_{max}, which was divided by the reference Npc (0.7) to calculate the maximum percentage of difference in Npc for statistically significant differences in feature abundance. For features with only one member, ΔNpc_{max} was calculated based on the minimum Npc at which that member was detected. This analysis was performed using the *in silico* metagenomes as well as seawater, human gut, freshwater, and peat soil metagenomes [24, 29, 30]. The approach was implemented in the Npc_max.R script, which can be used to calculate the ΔNpc_{max} of custom user data.

Metagenome quality filtering, assembly, and binning

A total of 85 marine metagenomes representing surface to 200 meters deep previously published by Hawley and colleagues [31] were downloaded from NCBI using the prefetch and fastq-dump (–-minReadLen 50) tools of the SRA toolkit. Raw reads were quality-filtered using bbduk.sh v38.18 (qtrim = w,3 trimq = 17 minlength = 70 tbo tossjunk = t cardinalityout = t). Reformat.sh v38.18 separated pair-end reads in two files and Nonpareil v3.4.1 [14] was employed to calculate metagenome coverage and diversity on cleaned forward reads (-T kmer -f fastq -X 50000 -t 8). Cleaned reads were assembled with SPAdes v3.15.5 (–meta –only-assembler -t 24 -k 21,33,55,77,99 127) [32] and contigs longer than 1 kb were selected for binning with MaxBin2 v2.2.7 [33] and metaBAT2 v2.15 [34]. Note that we assembled the original but not the subsampled metagenomes. Metagenome-assembled genomes (MAGs) were dereplicated with dRep v3.4.3 (–sa 0.95) [35] yielding a total of 219 species-like MAGs that were quality-assessed using CheckM v1.2.2 [36] and taxonomically classified with GTDB-tk v2.3.2 (r214) [37].

Antibiotic resistance genes relative abundance in wastewater metagenomes

To showcase the impact of Npc standardization in avoiding biased biological results due to differences in metagenome coverage, the dataset analyzed in Zhang *et al.*, 2021 [38] was reanalyzed here following the same pipeline they used but introducing standardization. Briefly, raw reads were downloaded with prefetch and fastq-dump and cleaned with fastp v0.21.0 (default parameters). Npc was calculated, and metagenomes subsampled to the same Npc. The abundance of antibiotic resistance genes (ARGs) in the original as well as the subsampled metagenomes was calculated with ARGs-OAP v3.2.4 [39] (default parameters) as in Zhang and colleagues. The estimateR function from the vegan R package [28]

was employed to calculate Chao1 index using as input the unnormalized counts for ARG subtypes provided by ARGs-OAP. Differential abundance test was performed with the t_test function of the R package rstatix [40] using the ARG subtypes abundance normalized by cell counts. Plots were drawn on R with ggplot2 v3.4.2 [26].

Results and discussion

Relative abundance and richness estimates can be affected by uneven diversity coverage

The motivation for this manuscript emerged from our own experience and mistakes. While analyzing metagenomes from complex peatland soils, we detected higher potential for oxygen respiration in the anoxic zone below the water table than in the surface oxic zone, which was apparently unexpected. Further investigation revealed differences in Npc between metagenomes as the source of this bias. This finding further highlighted the importance of addressing diversity coverage in metagenomic studies, motivating us to write this manuscript to raise awareness within the scientific community.

We first showcase the impact of differences in sequencing effort and Npc on derived conclusions by comparing the relative abundances of genomes in *in silico* metagenomes displaying different complexities in terms of evenness (high vs low evenness; we refer to these as even vs uneven metagenomes below, respectively), species richness (100 vs 1,000 distinct species), and microdiversity (1 vs 10 distinct genomes per species). The abundances of features in these metagenomes are thus known. The even and uneven datasets were produced using log-normal models of species abundances with different values of the location parameter (μ) and maximum/minimum abundance ratios (i.e. $\mu = 2$ and max/min ratio = 10,000 vs $\mu = 5$ and max/min ratio = 500, respectively). Note that high evenness does not mean equal abundances between species, which is rarely—if ever—observed in natural communities, but rather more even abundances compared to the uneven metagenomes. For assessing the impact of microdiversity, 10 genomes showing between 95% and 99.5% ANI among themselves were used. Nonpareil curves revealed evenness and species richness influence Npc estimation while microdiversity had a minimal impact (Fig. S1), presumably because the default parameters of Nonpareil are calibrated to assess redundancy at the 95% nucleotide identity (i.e. the species level). If intra-population (intraspecies) diversity is larger than this level, then such microdiversity would have a significant impact on Npc, and/or the default of 95% nucleotide identity of Nonpareil has to be adjusted accordingly. We did not explore this further as natural populations with greater microdiversity are rather uncommon [41]. Consequently, our subsequent analyses were limited to metagenomes exhibiting distinct levels of evenness and species richness.

The genome relative abundances were robustly estimated using read recruitments as SD normalized by GEQ (SD/GEQ, see Methods) in the complete metagenomes as well as in the subsampled metagenomes of varying Npc levels and number of sequences (Fig. 2A). Note that the reads were mapped back to the genomes used to simulate the metagenomes with no assembly step involved. The taxon rank curves of each metagenome type and replicate are shown in Fig. S2. The relative abundances of MAGs belonging to the same taxon (e.g. same order) were added up to represent the abundance of the taxon in each sample and the derived taxon abundances were directly compared between the complete metagenome and subsampled metagenomes at varying Npc levels. We consistently observed, across multiple taxa, increasing differences in their relative abundance in subsampled metagenomes of decreasing Npc compared to the original metagenome (Fig. 2B). For instance, the aggregated relative abundance of MAGs belonging to the order *Flavobacteriales* decreased by ~74% in metagenomes of Npc = 0.9 vs. 0.1. These results were initially counterintuitive because metagenomes, and their subsamples, are thought to be random subsets of the sequenced DNA; hence, relative abundance should be similar, if not identical, in subsamples. Further analysis revealed that many genomes were not detectable in subsampled metagenomes of low Npc coverage due to their relatively low abundance (Fig. 2C). Consequently, in the example of the *Flavobacteriales*, the drop of 74% in abundance in the completely sequenced metagenome vs a metagenome that was sequenced at Npc = 0.1 was simply due to diversity coverage differences, not actual differences in abundances. Collectively, these results indicated that sensitivity in detecting members of the feature, in this case a group of MAGs belonging to the same order, rather than the abundance of individual MAGs that make up the order, was responsible for the differences in the aggregated relative abundance of the feature observed as a function of the Npc and sequencing effort (Fig. 2B). In other words, the relative abundance of the individual MAGs remains stable in subsamples, as anticipated, but the number of detected MAGs increases with sequencing effort (or Npc coverage), which translates to higher aggregated abundance of the corresponding feature that the MAGs are assigned to.

Npc standardization mitigates errors associated with uneven metagenome coverage

The number of MAGs (or individual genes) detected can vary with metagenome coverage (Fig. 2C), which can significantly impact diversity and differential abundance analyses, leading to spurious or unreliable results. This bias can be minimized by standardizing to the same Npc, and this approach is advantageous compared to the common alternative of sequencing effort standardization, as we show below. Specifically, Even1000sp metagenomes contain 10 times more species than Even100sp metagenomes, a richness ratio accurately captured when standardizing to the same Npc but not when Npc was unequal, or when metagenomes were standardized to the same sequencing effort (Fig. 3). Furthermore, Npc standardization consistently produced the correct qualitative result, showing higher diversity for Even1000sp compared to Even100sp across sub-samplings. In contrast, sequencing effort standardization reversed this result at relatively low sequencing efforts, incorrectly showing higher diversity for Even100sp at 1 million reads, underscoring the limitations of read-based standardization. This strikingly misleading picture of the sequence effort standardization approach was due to the fact that the most abundant genomes in the Even100sp dataset were more abundant (and thus detectable) than the most abundant genomes in the Even1000sp dataset (see taxon rank curves in Fig. S2). We observed similar results with the same datasets when we performed alpha-diversity comparison using the Shannon and Simpson indexes (Fig. S3 and S4), further underscoring the robustness of Npc standardization. Comparable results were obtained based on the analyses of the Uneven1000sp and Uneven100sp datasets, although richness differences between the two datasets became inaccurate at low Npc values (<0.4) (Fig. S5). These findings highlight the strong qualitative (which dataset is higher/lower) and quantitative (how much higher/lower) benefits of standardizing to the same Npc for diversity comparisons.

To further quantify the importance of Npc standardization, we used the aggregated relative abundance at the phylum level

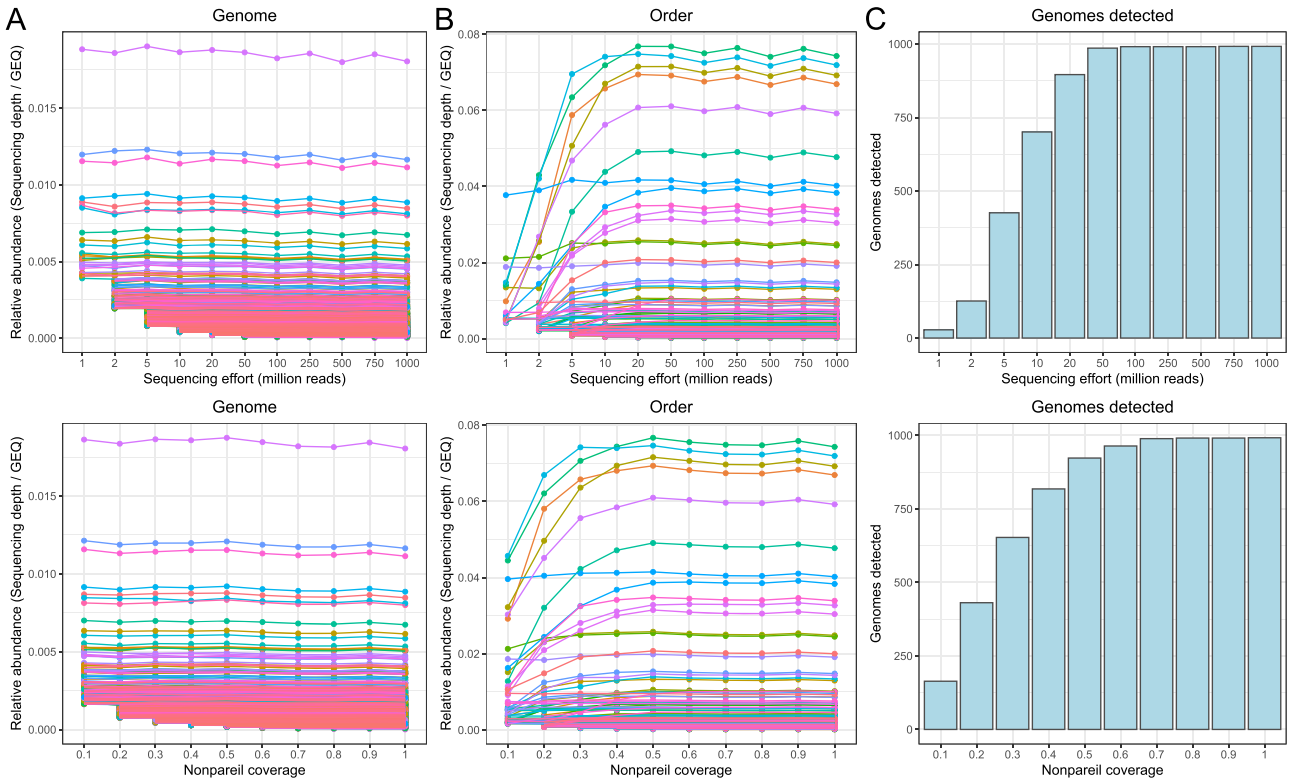


Figure 2. Impact of sequencing effort (top) and Npc (bottom) on the relative abundance estimate of individual genomes (A), groups of genomes (B), and the total number of detected genomes (C). Data were derived from an *in silico* metagenome with 1,000 species and an uneven genome distribution (Uneven1000sp). Each line in (A) and (B) represents the relative abundance of an individual genome or the aggregated relative abundance of a group of genomes within the same order, respectively. Note that the relative abundances of individual genomes remain stable (A), while the relative abundances of orders increase with higher sequencing effort and Npc (B), driven by the detection of more of their member genomes (C). The same picture was obtained from metagenomes with different richness and evenness.

to perform differential abundance analyses on metagenomes subsampled to a range of Npc values and using the results of the original (not subsampled) metagenomes as reference (i.e. correct result). As expected, the fraction of accurate tests increased with increasing Npc for all six metagenome combinations (even vs. uneven, 100sp vs. 1000sp, and their combinations; Fig. 4). This result was further reinforced by Bray-Curtis distances, which showed increasing divergence in the phyla abundance results between subsampled metagenomes and the original (full) metagenomes as Npc decreased (Fig. S6). In addition, differential abundance tests with Npc standardized metagenomes showed higher accuracy compared to unequal Npc metagenomes. For example, the comparison between even and uneven metagenomes with 1000 species yielded an average accuracy of 73.7% for unequal Npc metagenomes that increased to 86.3% when Npc standardized metagenomes were used, representing a ~13% improvement in accuracy (Table 2). However, while Npc standardization minimizes inaccuracies, it does not eliminate them entirely, as the accuracy was in all cases below 100%, reinforcing the importance of reporting Npc to contextualize results.

Npc standardization improves aggregated abundance estimates in metagenomes from natural environments

To demonstrate that the impact of Npc on the aggregated relative abundance of a group of MAGs (e.g. an order) is not due to the design (e.g. species abundance curve) of our *in silico* metagenomes, we also examined metagenomes from natural environments. Specifically, we used the metagenomes from the ocean depth profile produced by Hawley and colleagues [31]. The aggregated relative abundance of a group of MAGs belonging to the order *Marinisomatales* was estimated based on the SD/GEQ metric in the complete metagenomes and their subsampled datasets (Fig. 5). A set of subsampled metagenomes standardized to the same Npc produced a peak in *Marinisomatales* abundance at 150 meters, in agreement with the full, high-coverage metagenomes (average Npc = 0.8; Fig. 5A), while the profile based on unequal Npc revealed the opposite trend as an effect of metagenome coverage, not actual biological/ecological differences (Fig. S7). These results were attributable to the increasing detection of MAGs with higher Npc (Fig. 5B) rather than to changes in the relative abundance of individual MAGs. In fact, the relative abundance of MAGs remained stable across subsampled datasets as expected, given the random nature of metagenomes and our approach to create subsampled datasets, until MAGs became nondetectable in low coverage metagenomes (Fig. 5C). We also observed that subsampled metagenomes with equal coverage at medium-to-high Npc (≥0.5) showed the same depth-profile trends (i.e. an abundance peak at 150 m) compared to the full metagenomes (average Npc = 0.8) while those at 0.4 Npc or below displayed divergent trends (i.e., abundance peak at depths other than 150 m; Fig. 5). This observation is in line with the 0.6 Npc threshold previously proposed to ensure adequate metagenome coverage and assembly of the sampled community and thus, meaningful biological comparisons between metagenomes [16].

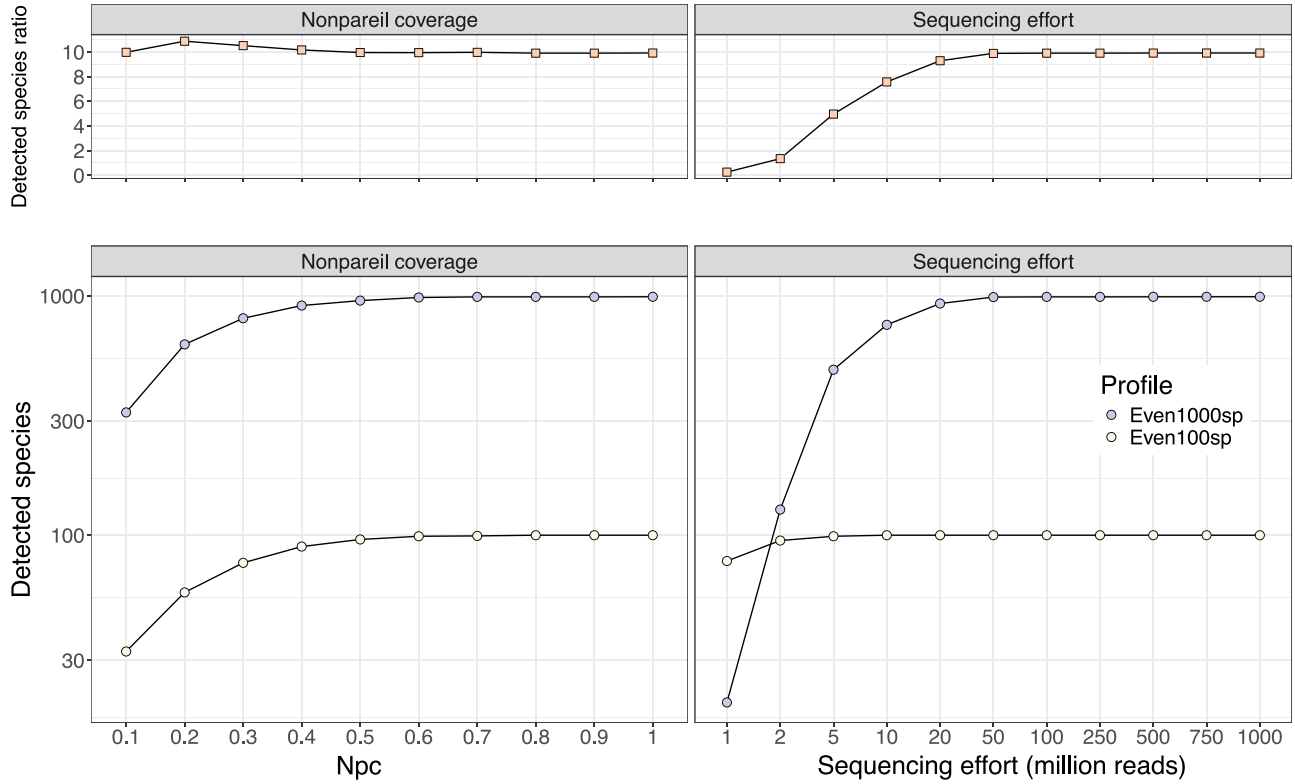


Figure 3. Number of detected species in metagenomes standardized to the same Npc (left) and sequencing effort (right). The real species ratio is 10 and the observed species ratio is shown on the top panels. Note that the Npc standardized metagenomes accurately capture the real richness differences between the metagenomes, contrasting with the sequencing effort standardization, which even provides the opposite result to the ground truth at relatively low sequence effort (1 million reads), showing higher diversity for the Even100sp than the Even1000sp datasets.

Table 2. Accuracy of differential abundance tests using metagenomes standardized to the same (equal Npc) or different Npc (unequal Npc). The underlying data is the same displayed in Fig. 4.

Metagenomes A		Metagenomes B			Average accuracy (%)		
# species	Evenness	# species	Evenness	Distinguishing factor	Unequal	Equal Npc	Improvement (equal Npc—unequal Npc; %)
1000	Even	1000	Uneven	Evenness	73.7	86.3	12.6
100	Even	100	Uneven		57.3	66.7	9.4
1000	Uneven	100	Uneven	Richness	89.0	95.1	6.0
1000	Even	100	Even		91.5	93.8	2.3
1000	Even	100	Uneven	Evenness and richness	92.9	95.1	2.2
100	Even	1000	Uneven		88.0	88.9	0.9

Published microbiome studies often report unreliable results

To address the impact of unequal Npc in published microbiome studies, we reanalyzed previously published richness and abundance results. For example, for features that involve genes, Zhang *et al.*, 2021 [38] apparently overestimated the ARG diversity found in effluent water samples of wastewater treatment plants (WWTP) due to the higher Npc of these samples compared to influent and sludge samples (Fig. S8A, B). Consistently, half of the ARGs subtypes identified as differentially abundant between different WWTPs previously [38] displayed a significantly different pattern when the metagenomes were subsampled to the same Npc (Fig. S8A, C).

Such cases are probably widespread in recent microbiome literature, unfortunately. Based on a literature search, the number of manuscripts performing differential abundance analysis with metagenomic data has increased from 89 manuscripts in 2014 to 2,880 in 2023 (Fig. S9). However, only 11 manuscripts out of the total 2,880 (0.38%) in 2023 considered Nonpareil in their analyses. Therefore, despite previous Nonpareil manuscripts having been cited over 600 times at the time of this writing, the tool has not been employed in the context of differential abundance analyses. In addition, popular statistical tools for differential feature abundance analysis of amplicon or transcriptomic data, such as metagenomeSeq, DESeq2, and edgeR [17, 42–44], apply different approaches to normalize read counts based on library size (e.g. CSS, RLE, and TMM) but do not incorporate differences in gene/genome length nor diversity coverage and make assumptions that are violated by metagenomic data (see Supplementary Note 1), limiting their usefulness for such data.

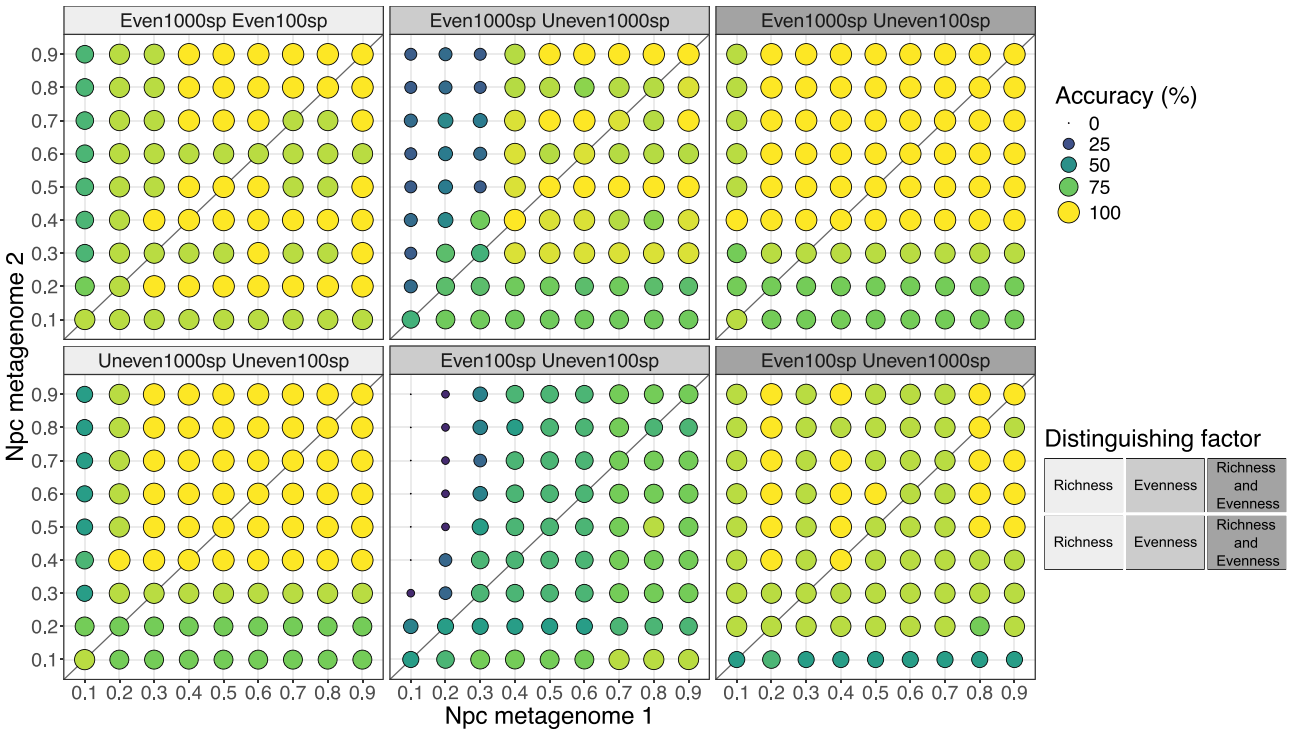


Figure 4. Accuracy of differential abundance tests across subsampled metagenomes of varying Npc. Aggregated relative abundances at the phylum level in the subsampled datasets were statistically compared between metagenome types. The significance of the test was compared to that obtained with the original metagenomes in order to assess the accuracy of the result (e.g. *Acidobacteriota* abundance higher in metagenome 1 than in metagenome 2). A total of nine phyla were compared, with the exception of Even1000sp vs. Uneven1000sp with 17 phyla. Each panel displays the result for one metagenome type pair (see panel titles). The diagonal line shows the comparison between metagenomes standardized to the same Npc. Abundances were compared based on a pairwise t-test among metagenome types with *P*-value adjusted using the Benjamini–Hochberg (BH) correction. The dot size and color indicate the percentage of tests performed that correctly detected the difference in abundance based on the complete metagenomes (i.e. accuracy; see figure key). Note that accuracy increases as Npc increases.

The need for Npc standardization depends upon the relative abundance of the target feature

We next opted to answer the question of how similar, in terms of Npc, the metagenomes should be in order to provide reliable relative abundance comparisons. For this, we calculated the maximum difference in Npc, as a fraction of the sample's Npc, which provided statistically identical results in terms of the relative abundance of a feature between the original and the subsampled metagenomes (ΔNpc_{max}; see Methods); i.e. the subsample provided unbiased results. The histogram of ΔNpc_{max} values showed a wide and even distribution with no predominant peaks (Fig. S10), suggesting that this parameter is highly variable, and there is no universal cutoff. Instead, we observed that the ΔNpc_{max} positively correlated with the average relative abundance of the feature of interest for the *in silico* metagenomes ($R^2 = 0.55–0.27$, *P*-value<.001, Fig. S11) as well as for various metagenomes from different environments ($R^2 = 0.57$, *P*-value<.001, Fig. S12), revealing no sample- or habitat-specific biases. These findings suggested that assessing abundant features, and thus robustly detected even in subsampled metagenomes, may not be biased when comparing metagenomes of different Npc while assessing low abundance features is more likely to be biased even when the ΔNpc is relatively low. For example, the estimate of abundance for a taxon with an abundance of 0.1% (measured as SD/GEQ∗100) of the total metagenome will not be biased unless ΔNpc is higher than 50%, based on the marine metagenomes with medium-to-high Npc used here. In contrast, the estimate for a taxon with an abundance of 0.02% will be biased if the ΔNpc is higher than 10%. The correlations shown on Fig. S12 can be useful to identify situations where this bias will be important to take into account or not.

Recommendations for comparative metagenomic analyses

Assessing features with single members (e.g. a MAG or a specific gene allele) is not problematic with respect to diversity coverage because metagenomes are random samples of diversity by nature. Hence, as long as the MAG/gene is confidently detected, relative abundance should be preserved in subsampled metagenomes. For features with multiple members (e.g. a bacterial genus or a gene family), sequencing high-coverage metagenomes (e.g. Npc > 0.9) is the most effective way to minimize metagenome coverage-associated biases. However, obtaining high metagenome coverage is not practical for several environments, such as soils, or studies with many samples. For such cases, data standardization is likely necessary for comparative purposes, but determining the optimal standardization approach in any given situation may be challenging. To assist researchers with this challenge, we provide a decision tree that can be used to guide comparative metagenomic analyses under different scenarios (Fig. 6). Specifically, caution is needed when the metagenome coverage is not high (e.g. Npc < 0.9), and there are differences between the Npc of the metagenomes being compared. In such cases, we recommend first calculating the ΔNpc_{max} of the feature of interest (a group of MAGs/genes representing a taxon/function) based on the metagenome(s) with the highest Npc in the comparison performed using the Npc_max.R script. If the ΔNpc_{max} is lower than the actual ΔNpc between the metagenomes compared, the

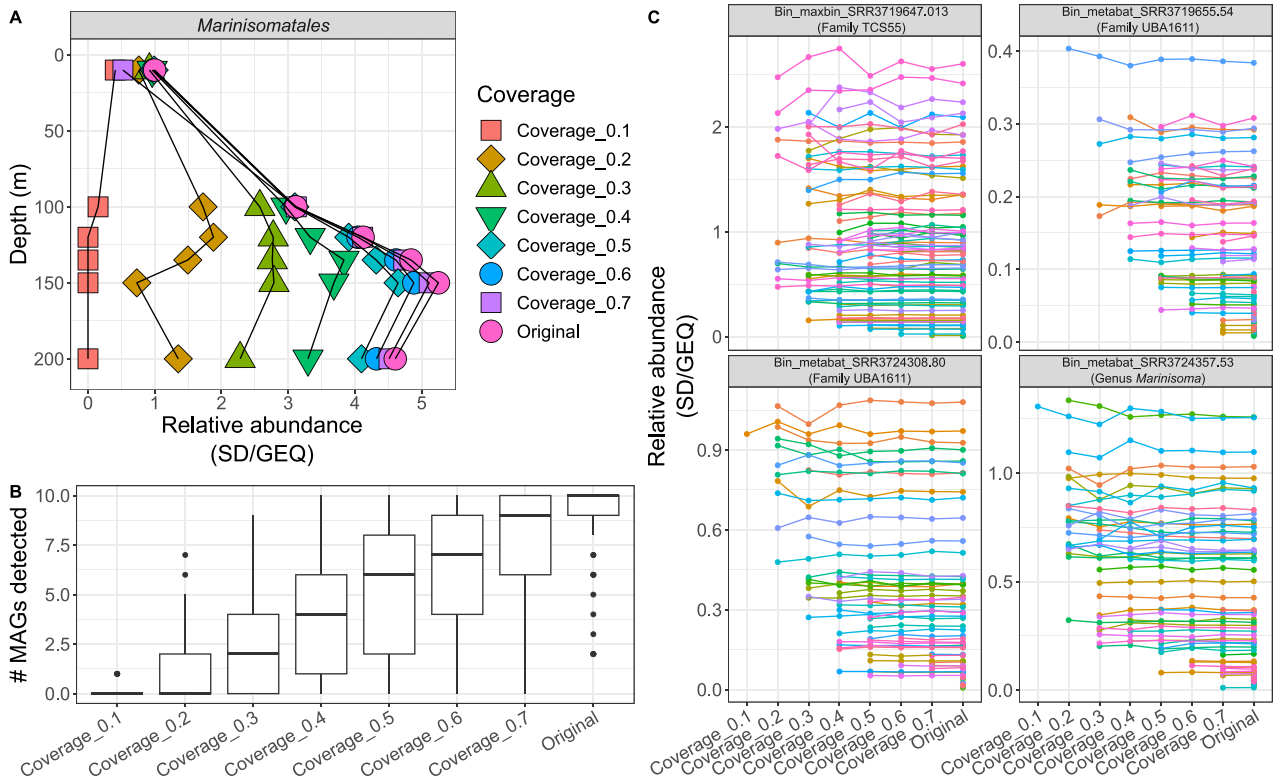


Figure 5. Effect of metagenome coverage on the aggregated relative abundance of a feature. (A) Aggregated relative abundance (x-axis), calculated as SD divided by GEQ, of a group of MAGs belonging to the order *Marinisomatales* is shown (x-axis) against the depth of the water column that the corresponding metagenomes were obtained from (y-axis). The average abundances from replicate metagenomes at each depth (10 m $n = 8$ replicates; 100 m $n = 15$; 120 m $n = 12$; 135 m $n = 11$; 150 m $n = 15$; 200 m $n = 16$) and subsampled metagenomes are shown. Note the increase in relative abundance as metagenome coverage increases. (B) Number of *Marinisomatales* MAGs (y-axis) detected in subsampled metagenomes (x-axis). Note the increase in the number of MAGs detected as metagenome coverage increases. (C) Relative abundance (y-axis) of individual MAGs belonging to *Marinisomatales* across subsampled metagenomes (x-axis). Each panel represents an individual MAG and each line and color indicate the relative abundance of such MAG across sub-samplings of the same sample. Note the consistency of relative abundance values across subsampled metagenomes but also that several MAGs become undetectable at low coverage subsamples (and thus, after this point, they do not contribute to the relative abundance of the *Marinisomatales* order).

comparison might be biased due to differences in metagenome coverage. In the latter case, we recommend calculating the relative abundance of the feature in metagenomes that have been standardized by subsampling to the same Npc (e.g. to the lowest Npc) (Fig. 6). Relative abundance and/or diversity values obtained after Npc standardization can then be used for more reliable comparative analyses. Metagenome subsampling and read mapping to calculate abundances in subsampled metagenome can be computationally expensive and requires experienced users. To assist with this step, we provide a fast, easy-to-use script (Npc_standardization.R) that accurately estimates relative abundance in subsampled metagenomes at a given, user-defined Npc using the abundance in the original complete metagenome. The abundance predictions of our tool showed perfect correlation ($R^2 = 1$) with minimal error (~1%) when compared to the observed abundance values obtained by actually subsampling and mapping the metagenomes reads (Fig. S13), and thus can be used for direct comparisons between metagenomes. Note that the Npc standardization approach presented here is useful for richness and differential abundance comparisons (i.e. macrodiversity). However, when the focus is on comparing differences in microdiversity within a species or population, standardization to the same SD of the target species/population should be employed [18]. Additionally, highly incomplete (e.g. <50% completeness) or contaminated (e.g. >10% contamination) MAGs could provide unreliable relative abundance estimates, as discussed previously [45]. Therefore, we recommend using only medium to high-quality MAGs as defined by MIMAG standards [46], for the types of comparative analyses described above.

It should be noted that even after Npc standardization (for metagenomes of uneven metagenome coverage) the relative abundances of features obtained could differ somewhat from the actual abundances in the completely sequenced samples (Figs 2 and 5). Thus, Npc standardization can make comparative analyses more robust but cannot guarantee the perfect accuracy of the resulting estimates, especially if standardizing to a relatively low coverage (e.g. Npc < 0.3; see Figs 4, S3, S4, S5 and S6), and the level of its success depends on the abundance distribution of the members of the feature being assessed (which determines at what sequencing effort/depth the features are detectable or not). The abundance distribution is typically unknown; therefore, dealing, in full, with this limitation is currently not feasible or requires more data (e.g. Npc > 0.9). Further, despite some controversy regarding the potential loss of statistical power with data subsampling [4, 9, 47], avoiding false positives commonly outweighs the risk of missing significant comparisons (false negatives), making rarefaction a pragmatic and advantageous approach compared to the alternative of analyzing the data as is. Moreover, since the abundance predictions generated by the Npc_standardization.R script produce maximum-likelihood estimates based on the complete data, the expected central value for abundances is adjusted to be comparable across datasets

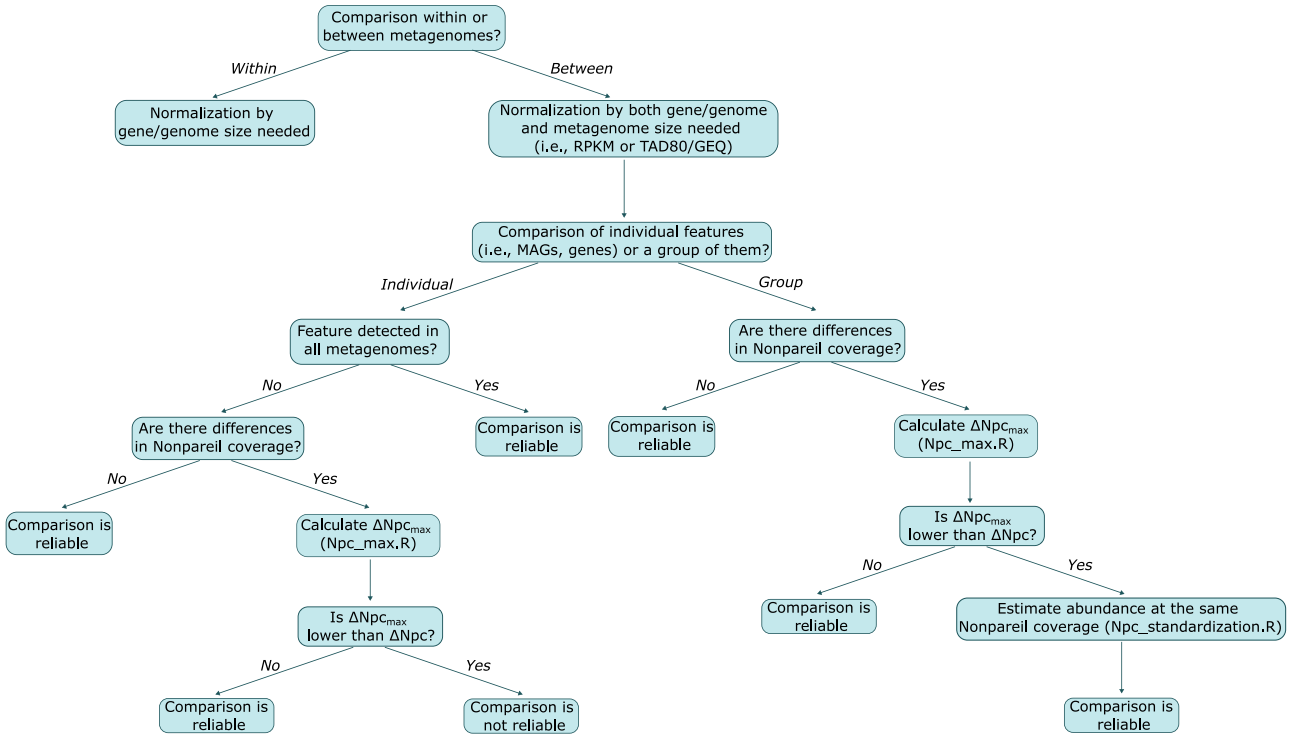


Figure 6. Decision tree to guide analysis of differential feature abundance or diversity between metagenomes.

without a corresponding increase in the estimation error and no data is discarded. Indeed, we observed ~2.1% average deviation within the observed SD values after 10 random subsampling of the same metagenome, which is in the same range or even higher than the ~1% average deviation of the same abundance value estimated by our script. Therefore, we recommend using this standardization approach and our method outlined above when possible, and resorting to random subsampling only for more complex cases not covered by our tool (e.g. when users want to assemble each subsample or create read recruitment plots).

Conclusion

In summary, we suggest to: (i) aim to sequence high coverage metagenomes to minimize the fraction of the community not covered, (ii) be aware of and calculate metagenome coverage, (iii) standardize the data to the same Npc when there are substantial differences in Npc (e.g. $\Delta Npc_{max} < \Delta Npc$), and (iv) be aware that results from differential abundance or diversity analyses for the obtained Npc may differ in higher coverage metagenomes. Accordingly, we expect that the recommendations provided will help to minimize biases in comparative metagenomic studies, thereby facilitating the generation of quantitative, standardized, and meaningful results.

Acknowledgements

We are grateful to Janet K. Hatt for her assistance in proofreading the manuscript.

Author contributions

B.A.R., L.M.R. and K.T.K. designed research; B.A.R. performed the analyses and developed the code; K.T.K. obtained fundings to support this work; B.A.R. and K.T.K. wrote the manuscript with the feedback from L.M.R.

Supplementary material

Supplementary material is available at *ISME Communications* online.

Conflicts of interest

None declared.

Funding

This work has been supported, in part, by the US Department of Energy (Award No DE-SC0023297) to K.T.K.

Data availability

The scripts and instructions to run them are available at https://github.com/baldeguer-riquelme/Nonpareil-coverage-standardization.

References

1. Willis AD. Rarefaction, alpha diversity, and statistics. *Front Microbiol* 2019;10. https://doi.org/10.3389/fmicb.2019.02407
2. Hughes JB, Hellmann JJ, Ricketts TH *et al.* Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* 2001;**67**:4399–406. https://doi.org/10.1128/AEM.67.10.4399-4406.2001
3. Good IJ. The population frequencies of species and the estimation of populations parameters. *Biometrika* 1953;**40**:237–64. https://doi.org/10.1093/biomet/40.3-4.237
4. Schloss PD. Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. *mSphere* 2024;**9**:e00354-23. https://doi.org/10.1128/msphere.00354-23

5. Weiss S, Xu ZZ, Peddada S *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 2017;**5**:27. https://doi.org/10.1186/s40168-017-0237-y
6. Cameron ES, Schmidt PJ, Tremblay BJ-M *et al.* Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Sci Rep* 2021;**11**:22302. https://doi.org/10.1038/s41598-021-01636-1
7. Bolyen E, Rideout JR, Dillon MR *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 2019;**37**:852–7. https://doi.org/10.1038/s41587-019-0209-9
8. Schloss PD, Westcott SL, Ryabin T *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;**75**:7537–41. https://doi.org/10.1128/AEM.01541-09
9. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;**10**:e1003531. https://doi.org/10.1371/journal.pcbi.1003531
10. Chao A, Jost L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 2012;**93**:2533–47. https://doi.org/10.1890/11-1952.1
11. Jost L. The relation between evenness and diversity. *Diversity (Basel)* 2010;**2**:207–32. https://doi.org/10.3390/d2020207
12. Roswell M, Dushoff J, Winfree R. A conceptual guide to measuring species diversity. *Oikos* 2021;**130**:321–38. https://doi.org/10.1111/oik.07202
13. Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity (hill numbers). *Methods Ecol Evol* 2016;**7**:1451–6. https://doi.org/10.1111/2041-210X.12613
14. Rodriguez-R LM, Konstantinidis KT. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 2014;**30**:629–35. https://doi.org/10.1093/bioinformatics/btt584
15. Rodriguez-R LM, Gunturu S, Tiedje JM *et al.* Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* 2018;3:10.1128/msystems.00039-18. https://doi.org/10.1128/mSystems.00039-18
16. Rodriguez-R LM, Konstantinidis KT. Estimating coverage in metagenomic data sets and why it matters. *ISME J* 2014;**8**:2349–51. https://doi.org/10.1038/ismej.2014.76
17. Pereira MB, Wallroth M, Jonsson V *et al.* Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 2018;**19**:274. https://doi.org/10.1186/s12864-018-4637-6
18. Bustos-Caparros E, Viver T, Gago J *et al.* Sequencing depth (coverage) can bias microbial intraspecies diversity estimates and how to account for it. *Research Square* 2024. https://doi.org/10.21203/rs.3.rs-4692663/v1
19. Manuel H. Mason – A Read Simulator for Second Generation Sequencing Data. *Technical Report FU Berlin* 2010.
20. Milhaven M, Pfeifer SP. Performance evaluation of six popular short-read simulators. *Heredity (Edinb)* 2023;**130**:55–63. https://doi.org/10.1038/s41437-022-00577-3
21. Shoemaker WR, Locey KJ, Lennon JT. A macroecological theory of microbial biodiversity. *Nat Ecol Evol* 2017;**1**:0107. https://doi.org/10.1038/s41559-017-0107
22. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015;**16**:51. https://doi.org/10.1186/s13059-015-0611-7
23. Castro JC, Rodriguez-R LM, Harvey WT *et al.* imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ* 2018;**6**:e5882. https://doi.org/10.7717/peerj.5882
24. Rodriguez-R LM, Tsementzi D, Luo C *et al.* Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. *Environ Microbiol* 2020;**22**:3394–412. https://doi.org/https://doi.org/10.1111/1462-2920.15112
25. Aroney STN, Newell RJP, Nissen J *et al. Woodcroft BJ*. CoverM: read alignment statistics for metagenomics *Bioinformatics* 2025;**41**:btaf147.
26. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016. https://doi.org/10.1007/978-3-319-24277-4
27. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 1988;**2**:231–9. https://doi.org/10.1016/0888-7543(88)90007-9
28. Oksanen J. *Multivariate Analysis of Ecological Communities in R: Vegan Tutorial*. 2011;**2011**. https://doi.org/10.1371/journal.pone.0024732
29. Kim M, Rodriguez-R LM, Hatt JK *et al.* Higher pathogen load in children from Mozambique vs. USA revealed by comparative fecal microbiome profiling. *ISME Communications* 2022;**2**:74. https://doi.org/10.1038/s43705-022-00154-z
30. Duchesneau K, Aldeguer Riquelme B, Petro C. *et al.* Northern peatland microbial communities exhibit resistance to warming and acquire electron acceptors from soil organic matter. *Nature Communications* 2025;**16**:6869. https://doi.org/10.1038/s41467-025-61664-7
31. Hawley AK, Torres-Beltrán M, Zaikova E *et al.* A compendium of multi-omic sequence information from the Saanich inlet water column. *Sci Data* 2017;**4**:170160. https://doi.org/10.1038/sdata.2017.160
32. Nurk S, Meleshko D, Korobeynikov A *et al.* MetaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34. https://doi.org/10.1101/GR.213959.116/-/DC1
33. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;**32**:605–7. https://doi.org/10.1093/bioinformatics/btv638
34. Kang DD, Li F, Kirton E *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. https://doi.org/10.7717/peerj.7359
35. Olm MR, Brown CT, Brooks B *et al.* DRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;**11**:2864–8. https://doi.org/10.1038/ismej.2017.126
36. Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55. https://doi.org/10.1101/gr.186072.114
37. Chaumeil PA, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 2020;**36**:1925–7. https://doi.org/10.1093/bioinformatics/btz848
38. Zhang D, Peng Y, Chan C-L *et al.* Metagenomic survey reveals more diverse and abundant antibiotic resistance genes in municipal wastewater than hospital wastewater. *Front Microbiol* 2021;12. https://doi.org/10.3389/fmicb.2021.712843
39. Yang Y, Jiang X, Chai B *et al.* ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic

data using an integrated structured ARG-database. *Bioinformatics* 2016;**32**:2346–51. https://doi.org/10.1093/bioinformatics/btw136

40. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. 2023. https://CRAN.R-project.org/package=rstatix, 2023.
41. Jain C, Rodriguez-R LM, Phillippy AM *et al.* High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;**9**:5114. https://doi.org/10.1038/s41467-018-07641-9
42. Paulson JN, Stine OC, Bravo HC *et al.* Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;**10**:1200–2. https://doi.org/10.1038/nmeth.2658
43. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data | bioinformatics | Oxford academic. *Bioinformatics* 2010;**26**:139–40. https://doi.org/10.1093/bioinformatics/btp616
44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-550-8
45. Meziti A, Rodriguez-R LM, Hatt JK *et al.* The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same Fecal sample. *Appl Environ Microbiol* 2021;**87**:e02593-20. https://doi.org/10.1128/AEM.02593-20
46. Bowers RM, Kyrpides NC, Stepanauskas R *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31. https://doi.org/10.1038/nbt.3893
47. Schloss PD. Waste not, want not: revisiting the analysis that called into question the practice of rarefaction. *mSphere* 2024;**10**:e1003531. https://doi.org/10.1128/msphere.00355-23

© The Author(s) 2025. Published by Oxford University Press on behalf of the International Society for Microbial Ecology. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
ISME Communications, 2025, **5(1)**, ycaf140
https://doi.org/10.1093/ismeco/ycaf140
Original Article