Long Reads Enable Accurate Estimates of Complexity of Metagenomes

Anton Bankevich[1(✉)] and Pavel Pevzner[2]

[1] Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, Saint Petersburg, Russia
anton.bankevich@gmail.com
[2] Department of Computer Science and Engineering, University of California at San Diego, La Jolla, CA, USA

Abstract. Although reduced microbiome diversity has been linked to various diseases, estimating the diversity of bacterial communities (the number and the total length of distinct genomes within a metagenome) remains an open problem in microbial ecology. We describe the first analysis of microbial diversity using long reads without any assumption on the frequencies of genomes within a metagenome (parametric methods) and without requiring a large database that covers the total diversity (non-parametric methods). The long read technologies provide new insights into the diversity of metagenomes by interrogating rare species that remained below the radar of previous approaches based on short reads. We present a novel approach for estimating the diversity of metagenomes based on joint analysis of short and long reads and benchmark it on various datasets. We estimate that genomes comprising a human gut metagenome have total length varying from 1.3 to 3.5 billion nucleotides, with genomes responsible for 50% of total abundance having total length varying from only 40 to 60 million nucleotides. In contrast, genomes comprising an aquifer sediment metagenome have more than two-orders of magnitude larger total length (≈840 billion nucleotides).

Keywords: Microbal diversity · Metagenomics · Rare spieces

1 Introduction

Locey and Lennon [29] recently estimated that Earth is home to as many as 1 trillion microbial species. In contrast, Schloss [42] demonstrated that, despite rapidly increasing sequencing efforts, the retrieval of 16S rRNA genes is approaching saturation. They argued that one-third of the bacterial diversity has already been discovered, implying that Earth is home to only millions of, rather than a trillion, bacterial species. This discrepancy and the emerged controversy [1,25,26,35,44,55] illustrate that the challenge of evaluating the bacterial diversity remains unsolved both at the global scale and at a single sample (local) level [51]. However, estimating the number of microbial species

B. J. Raphael (Ed.): RECOMB 2018, LNBI 10812, pp. 1–20, 2018.
https://doi.org/10.1007/978-3-319-89929-9_1

in a *given* sample is a more tractable yet difficult problem in microbial ecology [19,28,38,43]. Regrettably, many such estimates are inaccurate since most species in any metagenome belong to the rare biosphere [20,31,48]. Indeed, 16S rRNA libraries often capture only a small fraction of the sample diversity, resulting in large variations of the diversity estimates even for samples from similar environments. For example, the estimates of the number of microbial species in soil samples vary from hundreds [22], to tens of thousands [6,8] to a million [10].

Since the terms "diversity" and "complexity" have multiple interpretations in microbial ecology, we use the terms *metagenome richness* (the total number of species in a metagenome) and *metagenome capacity* (the total genome length of all species in a metagenome). Metagenome capacity can also be defined as metagenome richness multiplied by the estimated average length of genomes in the sample.

Since estimating the richness and capacity of metagenomes is a fundamental problem in microbial ecology, new approaches for solving this problem are needed. This challenge is further amplified by recent discoveries that linked the reduced diversity to various diseases. For example, reduced bacterial diversity results in an increased frequency of death in the allogeneic stem cell transplantation [50] and represents a biomarker for psoriatic arthritis [41]. On the other hand, increased bacterial diversity is associated with human papillomavirus infections [13] and White Plague Disease in corals [49].

The previous studies of bacterial diversity were primarily based on either *parametric* or *non-parametric* approaches [5,12,15,18,22,37]. Various parametric distributions were chosen to approximate the frequency distribution of captured species, and to project them to estimate how many more species must be present in the metagenome. This approach has been challenged since it is unclear which parametric distributions adequately model a given sample [17,54]. Furthermore, 16S rRNA data represents frequencies of specific PCR products that may differ in abundance relative to the corresponding bacterial species in a metagenome. In addition, they suffer from biases arising from various levels of primer matching in different species, inability to amplify taxa whose 16S rRNA differs from known ones, the variable number of 16S rRNA operons, and presence of highly diverged genomes with nearly identical 16S rRNAs [11,21,39].

As Hong et al. [17] discussed, applications of various probability distributions for evaluating the complexity of the metagenome have often been statistically incorrect. Moreover, even if some metagenomes follow a certain (e.g., exponential) distribution of frequencies, others may significantly deviate from this arbitrarily chosen model. For example, there is no reason to believe that the frequences of species for a soil metagenome and a human microbiome follow the same type of parametric probability distribution.

The alternative non-parametric estimators of microbial diversity [22] require a large database that covers all species in a sample, the condition that is typically violated. As a result, such estimates greatly underestimate the richness of metagenomes. As Lladser et al. [28] noted, most microbial communities have not been sufficiently deeply sampled yet to test the suitability of both parametric and non-parametric models.

With proliferation of metagenomics datasets covering entire genomes, it is important to have an independent way of estimating the richness of metagenomes that is not limited to sampling of 16S rRNA. Also, all previous studies attempted to estimate the metagenome richness/capacity using *short* reads that have limitations with respect to analyzing rare species. Since rare species within a metagenome typically have low coverage, they are hardly ever assembled into long contigs, making it difficult to estimate the richness, capacity, and the distribution of frequencies of various species within a metagenome. For example, bacterial species with coverage below 15X typically result in low-quality assemblies, making it difficult to estimate the number of such species or their total genome length. Since such rare species account for the lion's share of genomes in many metagenomes [20], they remain below the radar of modern sequencing technologies.

Below we describe the first analysis of microbial diversity using long reads without any assumption about the distribution of frequencies of genomes within a metagenome (parametric methods) and without requiring a database that covers the total diversity (non-parametric methods).

Our approach to estimating the capacity of a metagenome uses joint analysis of short reads (such as Illumina reads) with long reads (such as TruSeq Synthetic Long Reads, Pacific Biosciences reads, or Oxford Nanopores reads) and rests on a new insight based on *geometric probability* arguments rather than on merely applying the previously proposed approaches to SLRs. We view each long read as a "subgenome" and map all short reads to each subgenome to estimate its abundances. Afterwards, we apply geometric probability arguments to estimate the capacity of the entire metagenome from the abundances of all its subgenomes. We emphasize that our new approach requires *both* long and short reads (i.e., it does not work for short reads only or for long reads only) and demonstrate that it estimates the metagenome capacity and accounts for rare species even if their coverage by reads is below 0.01X, i.e., the species that are not "seen" by the state-of-the-art metagenome assemblers aimed at short reads. We apply our approach to estimate the complexity of the human gut metagenome (in a healthy individual and in multiple samples from a Crohn's disease patient at various stages of the disease) and in an aquifer sediment metagenome.

Although long reads are still rarely used for analyzing metagenomes, they have a potential to be widely used in future metagenomics projects when their cost reduces or when the *read until* technology [30] developed by Oxford Nanopores becomes widely available. We illustrate our approach using the TruSeq *Synthetic Long Reads (SLR)* technology that represents the first long read technology successfully applied to metagenomic studies [46]. The SLR technology generates accurate long virtual reads [27,32,52] that provided new insights into diversity of low abundance species in various metagenomes and revealed complex sub-partitioning of metagenomes into dozens of strains of the same species [23,46,53]. In addition, SLRs extend 16S rRNA studies, aimed at analyzing taxonomic diversity, by insights into functional diversity of rare species that often provide ecological impact through highly expressed transcriptomes and proteomes [20].

Although this paper focuses on SLRs, our method for estimating the capacity of a metagenome is applicable to long error-prone Single Molecule Sequencing (SMS) reads as well (see Appendix "Estimating metagenome capacity using long error prone SMS reads").

2 Results

Defining the Capacity of a Metagenome. Figure 1 (left) shows the histogram of frequencies of genomes comprising the artificial MOCK metagenome (with richness 20 and capacity 67 Mb) formed by mixing DNA from 20 isolate genomes [23]. However, the standard measures of richness and capacity depend on the taxonomic definition of a species and do not account for fragments shared by different species within a metagenome and for the fact that some species are represented by multiple similar strains. For example, if a metagenome contains two strains that share 99% of their genomes, should we count them as two genomes (and sum up their lengths) or as a single genome? Similarly, if a genome has a plasmid with multiplicity 100, should we count this plasmid 100 times towards the genome length or just one time?

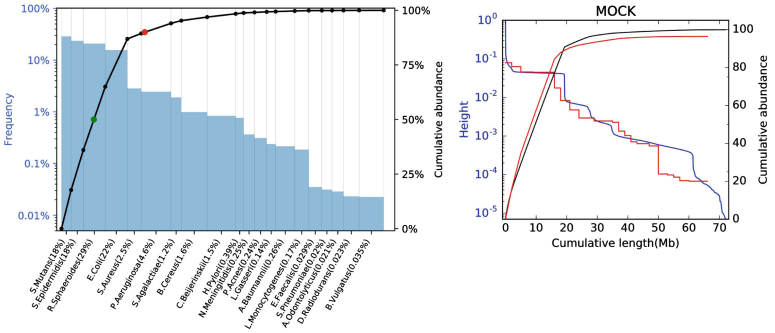

Fig. 1. Real (left) and estimated (right) real frequences of genomes and the abundance plot (left) and estimated frequency (height) and abundance plot (right) for the MOCK dataset formed by 20 bacterial species. (Left) frequences (in the logarithmic scale) of genomes in the MOCK dataset vary from ≈0.02% to ≈28%. The green and red points on the abundance plot correspond to M50 ≈ 7 Mb and M90 ≈ 17.5 Mb, respectively. The numbers next to the species names represent abundances (note that ranking based on frequencies differs from ranking based on abundances). The genome abundance is approximated as the fraction of short reads originating from this genome. The genome frequency is computed as the genome abundance normalized by the genome length. (Right) Blue and black curves show the estimated frequency (height) histogram and the abundance plot as predicted by our methods in comparison with red curves constructed using the known set of references. (Color figure online)

To address these issues, we introduce the concept of the *de Bruijn capacity* of a metagenome; this is motivated by a popular approach to genome assembly. We

construct the *de Bruijn graph* [7] of all genomes in the metagenome (including plasmids, viruses, unicellular eukaryotes, etc.), transform it into the *assembly graph* using SPAdes [2] (collapsing small variations in repeats), and define the de Bruijn capacity of a metagenome as the total length of edges (contigs) in the assembly graph. The de Bruijn capacity of the `MOCK` metagenome (58 Mb) is smaller than its total length (67 Mb) since this dataset contains a number of similar genomes and long intragenomic repeats.

Construction of the assembly graph is defined by a parameter *bubble length* that controls the percent identity level used for collapsing regions from various strains into a single contig in the assembly graph. For example, the default bubble length value in the SPAdes assembler [2] roughly corresponds to 98%–99% percent identity with respect to the taxomic definition of a strain (increasing the bubble length parameter will decrease the de Bruijn capacity). Thus, In the default setting, multiple strains that share 98%–99% of the genome (or multicopy plasmids) do not inflate the de Bruijn capacity. To reflect the stringency of taxonomic units in our estimator, one can vary the maximum number of mismatches and indels during alignment of short reads to SLRs.

We characterize each genome G in a metagenome M as a pair of numbers $(length(G), num(G))$, where $length(G)$ and $num(G)$ refer to the length and the copy number of this genome in the metagenome. We define *frequency* and *abundance* of a genome G as

$$frequency(G) = \frac{num(G)}{\sum_{G \in M} num(G)} \qquad abundance(G) = \frac{length(G) \cdot num(G)}{\sum_{G \in M} length(G) \cdot num(G)}$$

We note that the number of reads originating from a given genome within a metagenome is roughly proportional to its abundance rather than frequency (under the assumption of the uniform coverage).

The *frequency histogram* of a metagenome consisting of t genomes is defined by t bars with heights specified by the frequencies of the genomes and varying widths specified by the lengths of the genomes (Fig. 1 left). We define the *abundance plot* of a metagenome (Fig. 1 left) by considering t most frequent genomes within a metagenome and specifying a point $(length_t, abundance_t)$, where $length_t$ stands for the total length of these genomes and $abundance_t$ stands for the total abundance of these genomes (for each value of t). For each percentage x from 0% to 100%, we define the value $t(x)$ as the minimum t such that $abundance_t$ exceeds x%. In analogy to the Nx statistics for genome assembly [14], we define the *Mx statistics* for a metagenome as $length_{t(x)}$. For example, M50 $\approx$ 7 Mb and M90 $\approx$ 17.5 Mb for the `MOCK` dataset described below (Fig. 1, left).

Computing the abundance plots for complex microbial communities remains an open problem. Below we show how to construct such plots using the synergy between short and long reads.

The Total Rectangle Length Problem. We will first state an abstract probabilistic problem and later explain how it relates to the problem of estimating the

capacity of a metagenome. Consider a set M of rectangles, each rectangle R in this set characterized by its length $length(R)$ and height $height(R)$. We assume that the total area of rectangles is 1, i.e., $\sum_{R \in M} length(R) \cdot height(R) = 1$.

For a point ξ from one of the rectangles, we define $height(\xi)$ as the height of the rectangle that the point ξ falls into. We uniformly and independently sample N points from the total area of all rectangles and denote the rectangle the j-th point falls into as R_j, which is characterized by its length and height $(length_j, height_j)$ (Fig. 2). We further assume that the vector $(height_1, \ldots, height_N)$ is known but the vector $(length_1, \ldots, length_N)$ is unknown.

Let ξ be a random variable corresponding to the uniform sampling of a point from the total area of all rectangles. The described probabilistic process results in N samples of the random variable $height(\xi)$. Given a vector $(height_1, \ldots, height_N)$, our (somewhat ambitious) goal is to estimate the total length of all rectangles: $Length(M) = \sum_{R \in M} length(R)$.

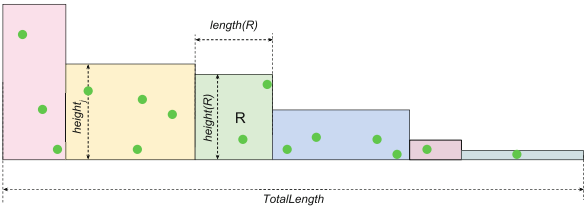

Fig. 2. The Total Rectangle Length Problem. N green points are sampled uniformly and independently from the probabilistic space defined by an (unknown) set of rectangles. Assuming that the heights of rectangles $(height_1, \ldots, height_N)$ these points fall into are known, the goal is to estimate the total length of all rectangles. (Color figure online)

The Total Rectangle Length Problem is intractable since the set of rectangles may contain a myriad of rectangles with extremely small heights and huge lengths whose total area is very small, e.g., significantly below $1/N$. Since these rectangles are unlikely to be sampled by any of N sampled points, our estimate cannot take into account their total length. We thus assume that all rectangles in the dataset have sufficiently large areas (e.g., larger than $1/N$) so that the probability of sampling each rectangle by at least one point is high.

Estimating the Total Length of Rectangles. Since the points are sampled uniformly, the probability of a point falling into a rectangle R equals the area of R: $area(R) = \mathbf{Pr}(\xi \in R)$. Thus, the length of the rectangle R is $length(R) = area(R)/height(R) = \mathbf{Pr}(\xi \in R)/height(R)$ and the total length of all rectangles can be estimated as:

$$Length(M) = \sum_{R \in M} \frac{\mathbf{Pr}(\xi \in R)}{height(R)} = \sum_{R \in M} \int_{x \in R} \frac{dx}{height(x)} = E\left(\frac{1}{height(\xi)}\right), \quad (1)$$

where E stands for the expectation of a random variable. Thus, by the law of large numbers:

$$Length(M) \approx \frac{1}{N} \sum_{j=1}^{N} \frac{1}{height_j}. \quad (2)$$

Moreover, according to the central limit theorem, the formula above is accurate, i.e., for large N, the variance of the estimate above is approximated as the variance of the random variable $\frac{1}{height(\xi)}$ divided by N. We will use this feature to estimate the metagenome capacity without any assumption on the parametric distribution of frequencies of genomes within a metagenome. See the Methods section for computing the abundance plot using a similar approach.

Representing a Metagenome as a Set of Rectangles. As before, we characterize each genome G in a metagenome M as a pair of numbers $(length(G), num(G))$ and define the total length of all genomes over all cells in a metagenomic sample as

$$sum = \sum_{G \in M} length(G) \cdot num(G)$$

The height of a genome G is defined as $height(G) = num(G)/sum$. Note that genome height is proportional to genome frequency. Thus each genome G is characterized by a rectangle $(length(G), height(G))$ and the total area of all rectangles is 1.

We assume that each genome within a metagenome results in a number of reads that is roughly proportional to its abundance. We also assume that each read is characterized by its starting position in one of the genomes and that these starting positions sample the genome uniformly and independently. Although the depth of coverage may deviate from the mean coverage in some genomic regions (e.g., in GC-rich regions or in the regions close to the origin of replication in the case of actively replicating genomes), such deviations are usually small. For the sake of simplicity, we assume that all genomes are circular, e.g., a read can start at any position of the genome so that there are no borderline artifacts (in the case of linear genomes, there are no reads that start within the last $i - 1$ positions of the genome, where i is the read length). Thus, starting positions model the random variable corresponding to the uniform sampling of a point from the total area of all rectangles.

We assume that N_{long} long reads and N_{short} short reads were sampled from the metagenome and that N_{short} is much larger than N_{long}. For example, various samples we analyzed contained ≈30–100 million short paired-end reads and ≈100–800 thousand long SLRs.

Each long defines a random point within the set of rectangles and our goal is to estimate the height of the rectangle this point falls into. We use short Illumina reads mapping to a given long read to estimate this height.

For simplicity, we assume that all short reads have the same length referred to as $|shortRead|$ (this condition holds for most sequencing projects). Since SLRs

are accurate, we assume that each short read aligned to an SLR *longRead* also maps to the corresponding genomic segment and vice versa. To avoid borderline effects, we assume that we can detect all short reads that align to *longRead*, even reads that start at the last position of *longRead*. To satisfy this condition, we shorten each SLR by $|shortRead|$ (or by the insert length) but map short reads using the enture span of each SLR. We define the number of short reads mapping to an SLR *longRead* as $number(longRead)$.

The fraction of short reads that map to *longRead* is expected to be approximately equal to the area in the rectangle space "above" *longRead*, i.e, to $|longRead| \cdot height(longRead)$, where $height(longRead)$ is defined as the height of the rectangle (genome) that contains *longRead*. Thus, the expected number of short reads that map to *longRead* (that we refer to as $E(number(longRead))$) can be estimated as

$$E(number(longRead)) = |longRead| \cdot height(longRead) \cdot N_{short} \tag{3}$$

Thus,

$$height(longRead) \approx \frac{number(longRead)}{N_{short} \cdot |longRead|}. \tag{4}$$

We just reduced the problem of estimating the capacity of a metagenome to the Total Rectangle Length Problem. We are given a set of N_{long} points (SLRs) that represent a uniform and independent sampling of an unknown set of rectangles (the metagenome). Each SLR is characterized by its length $|longRead_j|$ and the number of short reads $number_j$ mapping to this SLR (for $1 \le j \le N_{long}$). We estimate the height h_j of the j-th SLR read using formula 4. Thus, the capacity of the metagenome is estimated as

$$Capacity(Metagenome) \approx \frac{1}{N_{long}} \sum_{j=1}^{N_{long}} \frac{1}{height_j} \approx \frac{N_{short}}{N_{long}} \sum_{j=1}^{N_{long}} \frac{|longRead_j|}{number_j}. \tag{5}$$

In the Methods section we describe similar formulas for constructing the frequency histogram and the abundance plots. Formula 5 has limitations in analyzing extremely rare species, e.g., species that did not result in any SLRs or species that resulted in SLRs with extremely small coverage by short reads. Note that this formula was derived in two steps: estimation of the expectation of the inverse height (formula 2) and estimation of the SLR height through its coverage by short reads (formula 4). We discuss how to address the limitations of these steps in the Methods section.

Datasets. We analyzed the following metagenomics datasets based on SPAdes [2] and truSPAdes [3] assemblies of SLRs (see Appendix "TruSPAdes assemblies of MOCK, GUT, and SEDI datasets"):

The `SYNTH` synthetic community dataset is formed by a set of short Illumina reads from the genomic DNA mixture of 64 diverse bacterial and archaeal species (Shakya et al. [45]; SRA acc. no. SRX200676) that was used for benchmarking

the Omega assembler [16]. It contains ≈109 million Illumina HiSeq 100 bp paired-end reads with mean insert size of 206 bp. Since the reference genomes for all 64 species forming the SYNTH dataset are known, we used them to evaluate the accuracy of our estimator. The total length of the reference genomes for this dataset is ≈200 Mb and its de Bruijn complexity is ≈190 Mb.

The SYNTH dataset contains short Illumina reads but does not contain SLRs. We thus simulated 6306 virtual SLRs (providing the average coverage of 0.25 for the metagenome) for the SYNTH dataset by randomly selecting a short read, mapping it to one of the reference genomes, and extending the region covered by this read by N nucleotides in both directions (N was uniformly distributed between 1500 to 5500). This simulation protocol ensures that simulated SLRs are sampled from metagenome with the same probability distribution as the short reads.

The MOCK synthetic community dataset is formed by short Illumina reads and SLRs from the genomic DNA mixture of 20 bacterial species [23]. It contains ≈31 million Illumina paired-end short reads with mean insert size of 247 bp and ≈221 thousand SLR reads longer than 6 kb constructed from three sets of 384 barcoded read pools each. Since the reference genomes for all species forming the MOCK dataset are known, we used them to assess the accuracy of our estimator. The total length of the reference genomes for this dataset is ≈67 Mb and its de Bruijn complexity is ≈58 Mb.

The GUT dataset is formed from short Illumina reads and SLRs sampled from the gut microbiome of a healthy human male that was analyzed in Kuleshov et al. [23]. It contains ≈80 million paired-end short reads with mean insert size of 208 bp and seven sets of barcoded read pools (384 pools in each set) that resulted in ≈501 thousand SLR contigs longer than 6 kb. Using this dataset we provide a new estimate of the capacity of the human gut metagenome.

The SEDI dataset is formed from short Illumina reads and SLRs sampled from an aquifer sediment that was analyzed in Sharon et al. [46]. It contains ≈27 million paired-end short reads with mean insert size of 351 bp and three sets of barcoded read pools (384 pools in each set) that resulted in ≈215 thousand SLRs longer than 6 kb. Sharon et al. [46] revealed a high diversity of strains in the genomes of this dataset. We confirm findings of Sharon et al. [46] and turn their initial observation into an estimate of the SEDI metagenome capacity.

In difference from the SYNTH and MOCK datasets, the metagenome capacity of GUT and SEDI datasets remains unknown. In addition to these datasets, we analyzed a larger synthetic dataset and four human microbiome datasets from a patient suffering from the Crohn's disease (see Appendix).

Benchmarking. For each dataset, we estimated the capacity of the corresponding metagenome (using formula 2) and constructed the frequency histogram and the abundance plot (Fig. 1 (right) for MOCK dataset and Fig. 3 for the other three datasets) using formula 7. We analyzed 1000 SLRs with the highest coverage in each dataset (contributing to a small "bump" in the beginning of the frequency histograms) and confirmed that most of them arise from plasmids and

16S rRNAs. This finding suggests that highly-covered SLRs can be used for *de novo* assembly of new plasmids and characterization of previously unknown 16S rRNAs directly from metagenomics datasets. Recent attempts to address these problems using short reads with tools like RECYCLER [40] and PhylOTU [47] faced computational challenges since it remains unclear how to extract plasmids and 16 rRNAs from the complex de Bruijn graphs of metagenomes. In order to evaluate how our estimator deteriorates with reduction in coverage by short reads and/or long reads, we downsampled the entire datasets of short reads and SLRs (Table 1).

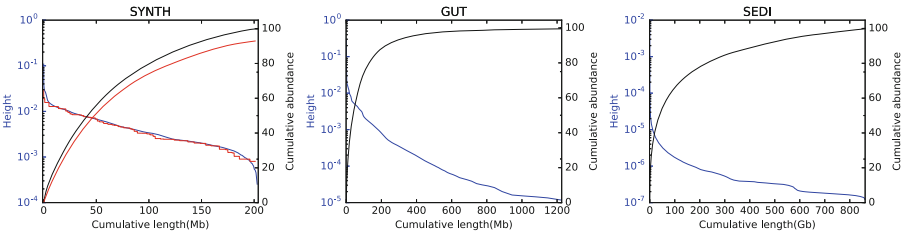

Fig. 3. Estimated frequency histograms (blue curve) and abundance plots (black curve) for `SYNTH`, `GUT`, and `SEDI` datasets. The distribution of heights (frequencies) of individual genomes within a metagenome was obtained based on alignments of short reads to SLRs. For the `SYNTH` dataset, we compared the constructed frequency histogram and abundance plot with the red plot representing the reference genomes with known abundancies. The y-axis of frequency histograms show the histogram of heights SLRs (in the decreasing order of heights) multiplied by 10^6, i.e., the probability that a random read falls into a 1 Mb long segment of the metagenome specified by the x coordinate. For the `GUT` dataset, M50 = 40 Mb and M90 = 230 Mb. For the `SEDI` dataset, M50 = 39 Gb and M90 = 432 Gb. (Color figure online)

`SYNTH`. As Table 1 illustrates, our estimator is accurate even with a small number of SLRs and short reads; e.g., even for short reads downsampled at 5%, deviation from the total metagenome size does not exceed 15%.

Note that the coverage of some genomes in the `SYNTH` dataset is as low as 6X [16]. Our capacity estimate remains accurate even at 0.1% downsampling, corresponding to the coverage by short reads for some genomes as low as 0.006. Note that the estimated capacity is accurate when almost all SLRs are covered by at least one short read.

`MOCK`. Table 1 illustrates that our formula accurately estimates the metagenome capacity when at least 7% of short reads are used. Note that while the `MOCK` dataset is subject to various biases that affect sampling of SLR and short reads (e.g., the GC bias), our formula is still accurate. Table 1 also shows that our formula generates stable capacity estimates even with a highly variable number of downsampled SLRs and suggests that there is a number of rare species in this metagenome.

Table 1. The metagenome complexity estimation (in Mb) for SYNTH and MOCK datasets. Columns correspond to downsampling of SLR reads, while rows correspond to downsampling of short reads. The last column shows the percentage of SLRs that were not covered by any reads from the downsampled set of short reads. Estimated metagenome capacity (in Mb) for SYNTH and MOCK datasets is 200 Mb and 67 Mb, respectively. Estimated de Bruijn capacity (in Mb) for for SYNTH and MOCK datasets is 190 Mb and 58 Mb, respectively.

SYNTH					
	Estimated metagenome capacity using				Fraction of uncovered SLRs
	100 SLRs	500 SLRs	2000 SLRs	10000 SLRs	
0.02%	156	144	147	150	51%
0.1%	204	220	209	218	28%
1%	194	241	224	230	0.4%
5%	182	221	203	205	0%
25%	179	212	198	201	0%
100%	179	209	196	199	0%
MOCK					
	Estimated metagenome capacity using				Fraction of uncovered SLRs
	100 SLRs	1000 SLRs	10000 SLRs	220748 SLRs	
1%	34	31	37	37	3%
7%	53	56	59	58	0.7%
20%	76	73	73	71	0.08%
100%	69	60	68	72	0.005%

GUT. We estimated the capacity of the human gut metagenome at ≈1.3 billion nucleotides, in line with previous estimates of the human gut microbiome richness [36]. Also, a rather small fraction of SLRs were not covered by reads (0.8%), suggesting that our estimate is accurate. Note that assembly of this dataset performed in Kuleshov et al. [23] resulted in contigs with total length of 656 Mb. Thus, the assembled contigs in the GUT dataset represent a large fraction of this metagenome.

SEDI. Our formula resulted in ≈840 Gb estimate for the capacity of this metagenome but ≈47% of SLRs were not covered by any short reads, suggesting that this metagenome is very diverse and that it contains a very large number of extremely rare species (with coverage 0.01X and below) which account for most of the total DNA in this metagenome. Thus, our formula is likely to underestimate the complexity of this metagenome. Note that the total length of assembled contigs for the SEDI dataset (204 Mb for contigs longer than 1 kb) is significantly lower that the estimated capacity of the metagenome. Since the large SEDI metagenome may include unicellular eukaryotes with large genomes (that are common in sediments [4]) and is likely to include a large fraction of relic DNA [26], it is difficult to estimate its richness.

3 Methods

Estimating the Abundance Plot. Let D be a value range of a random variable ξ with density p with respect to a measure μ. By considering $p(\xi)$ as a random variable, we have:

$$E\left(\frac{1}{p(\xi)}\right) = \int_D \frac{1}{p(x)} p(x) = \int_D 1 = |D|. \quad (6)$$

Thus, formula 1 is a special case of a more general formula for the value range size estimation. This interpretation also allows us to estimate the value of $|D_t|$, where $D_t = \{x \in D | p(x) < t\}$:

$$|D_t| = \int_{D_t} 1 = \int_{D_t} \frac{1}{p(x)} p(x) = \int_D \frac{1}{p(x)} \delta_{p(x)<t} p(x) = E\left(\frac{1}{p(\xi)} \cdot \delta_{p(\xi)<t}\right). \quad (7)$$

The right part of this formula can be estimated similarly to formula 2, resulting in the estimate of the frequency histogram. The graph of $|D_t|$ as a function of t gives the abundance plot of a metagenome. In practice estimation of frequency histogram can be constructed using the following method. Given the heights of SLRs in the decreasing order $(h_1, \ldots, h_{N_{long}})$ computed using formula 4, the frequency histogram consists of N_{long} bars with the j-th bar in the histogram having height h_j and width $\frac{1}{h_j}$. The abundance plot is merely the integral of the frequency histogram.

Variance of the Metagenome Capacity Estimator. We used the central limit theorem (CLT) as the basis of our estimator. The accuracy of the resulting formula in the CLT is defined by the variance of the random variable in question. For example, in the case when a significant fraction of a metagenome results in rectangles with extremely low height (e.g., rectangles with area less than $1/N_{long}$), the variance of the random variable is very high. We thus make an assumption that nearly entire metagenome is comprised from the genomes with sufficiently large frequencies to be captured by SLRs. Since typical SLR projects result in $10^5 - 10^6$ SLRs, this constraint implies that the metagenome that we are able to analyze mostly consists of species with frequencies exceeding 0.001%. Under this assumption, we can use the CLT to compute the variance of our estimator.

Accuracy of the Inverse Height Estimator. Formula 3 leads to an unbiased estimate of the SLR height $height(longRead)$ (given by formula 4). However, the value that we actually need to estimate is $\frac{1}{height(longRead)}$ and this estimation, given by formula 5, becomes biased. Below we analyze how this bias affects our estimation of the metagenome capacity.

We first consider a simple case when the metagenome consists of a single genome *Genome* and when SLRs sampled from *Genome* have the same length.

We also assume that the number of reads mapped to a genome fragment (and an SLR) follows the Poisson distribution:

$$number(longRead) \sim Poisson(\lambda), \tag{8}$$

where λ represents the expectation of the number of reads mapped to *longRead*. The value λ can be estimated as: $\lambda = |longRead| \cdot height(longRead) \cdot N_{short}$. We can now compute the value $|Genome|^*$, the genome length that is (erroneously) estimated by formula 5 instead of $|Genome|$:

$$|Genome|^* = N_{short} \cdot E\left(\frac{|longRead|}{Poisson(\lambda)} | Poisson(\lambda) \neq 0\right) \tag{9}$$

Note that since $height(Genome) \cdot |Genome| = 1$, the function δ, defined as $\frac{|Genome|^*}{|Genome|}$, depends only on the value of λ:

$$\begin{aligned}\delta(\lambda) &= \frac{|Genome|^*}{|Genome|} = \lambda \cdot E\left(\frac{1}{Poisson(\lambda)} | Poisson(\lambda) \neq 0\right) \\ &= \frac{\lambda}{1-e^{-\lambda}} \sum_{n=1}^{\infty} [(1/n) \cdot e^{-\lambda} \cdot \lambda^n / n!] = \frac{\lambda \cdot e^{-\lambda}}{1-e^{-\lambda}} (-\gamma - \ln(\lambda) - Ei(-\lambda))\end{aligned}$$

where $\gamma \approx 0.57721566$ is the *Euler-Mascheroni* constant and Ei is the *exponential integral* $Ei(z) = -\int_{-z}^{\infty} e^{-t} t^{-1} dt$. Thus, the expectation of the relative error in formula 5 is defined by $\delta(\lambda)$. The higher is the value of λ (which refers to the average number of short reads mapped to a long read), the closer δ is to 1. For example, if the expected number of short reads aligned to an SLR exceeds 15, the relative error of our estimate is at most 10%. Coverage of a typical 10 kb long SLR by 15 reads corresponds to genome coverage of $15 \cdot |shortRead|/|longRead| =$ 0.15X for short reads of length 100.

This analysis illustrates why long reads provide a much "deeper" look into the capacity of a metagenome than short reads. Indeed, it enables analysis of genomes with the coverage 0.15X and below as compared to the coverage 15X that is typically needed for assembling a genome within a metagenome from short reads. For genomes with a value of λ significantly less than 1, it turns out that most SLRs sampled from them have zero coverage by short reads. Thus, genomes with very low coverage contribute little to the estimate of the metagenome capacity.

4 Discussion

The recent bacterial census update [42] highlighted that high-throughput sequencing is based on short reads, while a high-quality census requires a high-throughput *full-length* 16S rRNA sequencing (rather than conventional short reads sequencing). It also illustrated the need for alternative technologies to analyze bacterial diversity such as single cell sequencing [21]. However, without

prior sorting, single cell sequencing mostly reports the abundant species. In contrast, a large fraction of individual genomes assembled from metagenomes had not been sequenced before [34]. However, the number of genomes reliably recovered from a metagenome is usually limited to hundreds at best, a small fraction of the total diversity of a metagenome. These difficulties highlight the need for a yet another technology for evaluating bacterial diversity. We showed that a combination of short-read and long-read sequencing technologies solves this problem even though each of these technologies separatily does not provide accurate estimates of the metagenome capacity. Although our analysis may be hampered by a potential metagenome sampling bias between short and long reads, our estimator of a metagenome complexity results in a useful approximation of the metagenome size.

Analysis of various metagenomics samples revealed that, although there often exists a small number of abundant species, thousands of low-abundance highly-diverged species account for most of the observed diversity. While this rare biosphere represents a source of genomic innovation [20], previous metagenomics studies, plagued by limitations of short reads technologies, were unable to evaluate its diversity. This study is the first attempt to estimate the diversity of the rare biosphere using a combination of short and long reads. Our analysis of the `SEDI` dataset illustrates, this rare biosphere may contain hundreds of thousands species even for a single soil sample. As the existing estimates of richness of soil and sediment bacterial communities differ by orders of magnitides, it would be interesting to apply our approach to analyzing other soil/sediment hybrid datasets when they become available.

Our approach also revealed significant variations in the diversity of the human gut metagenome in the case of an individual with the Crohn's disease. We envision that the metagenomics studies will soon move to generating a nearly complete census of all bacteria within microbiomes across the entire human population [33]. Our method will provide an estimate of the still unknown fraction of metagenomes that will be important for building such a census.

Acknowledgements. We are indebted to Chris Dupont, Rob Knight, and Glenn Tesler for providing numerous comments. Glenn Tesler also suggested using exponential integrals for analyzing the bias of our estimator. We are grateful to Yana Safonova, Andrey Bzikadse, Sergey Bankevich, Sergey Nurk, Alon Orlitsky, Ivan Tolstoganov, and Aleksandr Shlemov for many helpful discussions and help with preparation of this paper. This study was funded by the Russian Science Foundation (award 14-50-00069) and by the National Science Foundation (MCB-BSF award 1715911).

Appendix

TruSPAdes Assemblies of MOCK, GUT, and SEDI Datasets. The TruSeq SLR technology generates accurate and long virtual reads derived from pools of short reads [27,32,52]. It is based on fragmenting genomic DNA into large segments (≈10 kb long) and forming random pools of the resulting segments (each pool contains ≈300 segments). Next, these fragments are amplified,

sheared, and marked with a barcode that is unique to the pool. Afterwards, they are sequenced using the standard Illumina short reads technology. All short reads originating from the same barcode are assembled together resulting in a set of long contigs (this step is called the *SLR barcode assembly*). Ideally, the result of such sequencing effort for a single barcode is the collection of approximately 300 fragments (each fragment is ≈10 kb long) from a genome forming 300 long virtual reads. SLRs have low mismatch rate (about 0.1%), extremely low indel rate, and few misassemblies [3].

Table 2 presents results of barcode assembly of `MOCK`, `GUT` and `SEDI` datasets with truSPAdes.

Table 2. Results of truSPAdes assemblies of `MOCK`, `GUT` and `SEDI` datasets. Long SLRs are defined as SLRs longer than 6 kb.

	`MOCK`	`GUT`	`SEDI`
#SLRs	451036	1226918	210495
#long SLRs	220778	772833	157336
N50	9180	8625	8266
Avg. #long SLRs per barcode	191	287	136
Total length of SLRs (Gb)	2.9	8.4	1.5
Total length of long SLRs (Gb)	2.1	5.8	1.3

Analyzing the CAMI and CROHN Datasets. In addition to datasets described in the main text, we also analyzed a larger synthetic dataset and four human microbiome datasets from a patient suffering from the Crohn's disease.

The `CAMI` dataset is a simulated dataset generated by the "Critical Assessment of Metagenome Interpretation" (CAMI) initiative aimed at evaluating various approaches to analyzing metagenomes (http://www.cami-challenge.org/). We used a CAMI dataset simulated from 225 genomes and containing 150 million 100bp paired-end reads with mean insert size of 180bp (the errors in simulated reads are modelled after Illumina HiSeq). We simulated 50 thousand SLRs in the same way as for the `SYNTH` dataset. The total length of the reference genomes for this dataset is ≈820 Mb and its de Bruijn complexity is ≈770 Mb. Figure 4 shows that our estimator works well for the `CAMI` dataset.

The `CROHN` datasets are four human gut microbiome datasets from a patient with Crohn's disease. These datasets (`CROHN1`, `CROHN2`, `CROHN3`, `CROHN4`) represent a metagenomics time series collected at 12-28-2011, 04-29-2013, 11-16-2014 and 06-29-2015, respectively. Each of these datasets includes one Illumina paired-end library and one SLR library. Number of short reads in these datasets ranges from 150 to 230 millions with mean insert size ≈400 bp for all datasets. The number of SLRs ranges from 17 to 50 thousand. Assembly efforts for these datasets resulted in contigs of length 242, 172, 225 and 275 Mb for `CROHN1`, `CROHN2`, `CROHN3`, and `CROHN4` datasets respectively.

We estimated metagenome capacity for CROHN1, CROHN2, CROHN3, and CROHN4 datasets as 3.5, 2.0, 2.4, and 3.2 Gb, respectively. Values of M50 were estimated as 41, 61, 25, and 45 Mb, respectively, while values of M90 were estimated as 230, 490, 240, 250 Mb respectively. These estimates reveal large variations in metagenome capacity during the course of disease that go well beyond what can be estimated using short read assemblies. Correlation between metagenome capacity and antibiotics treatments for this metagenomics time series will be discussed elsewhere.

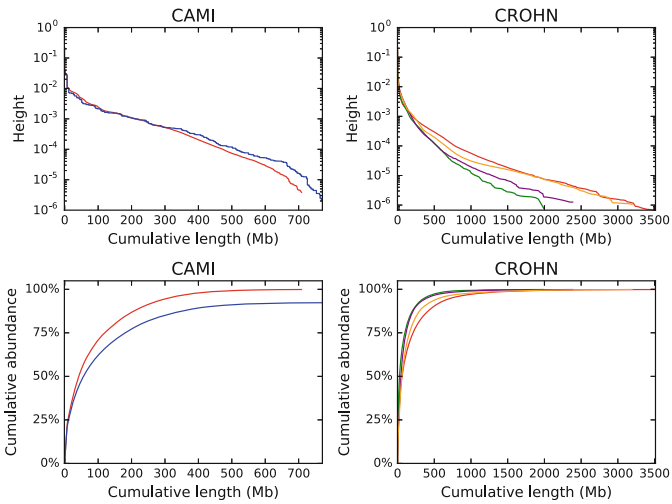

Fig. 4. Estimated frequency histograms and abundance plots for CAMI (left) and CROHN1, CROHN2, CROHN3, CROHN4 datasets (right). The distribution of heights (frequencies) of individual genomes within a metagenome was obtained based on alignments of short reads to SLRs. For the CAMI dataset, we compared the constructed plots with the blue plot representing the reference genomes with known abundancies.

Estimating Metagenome Capacity Using Long Error Prone SMS Reads. Although SMS reads (e.g., reads generated using Pacific Biosciences and Oxford Nanopores technologies) are still rarely used for analyzing metagenomes [9], they have a potential to be widely used in future metagenomics projects when their cost reduces and when the *read until* technology [30] developed by Oxford Nanopores becomes widely available. Below we show how to extend our approach for estimating the metagenome complexity using SMS reads.

SMS reads present an attractive alternative to TSLRs since their average length is higher and since they feature a uniform coverage depth that is not affected by the GC content. However, alignment of short Illumina reads against error-prone SMS reads is a more challenging task than their alignment against accurate TSLRs. We addressed this complication using the bowtie2 alignment tool [24] with specially selected parameters aimed at alignment of short Illumina

reads against error-prone SMS reads (-D 40 -R 3 -N 0 -L 17 -i S,1,0.50 –rdg 1,3 –rfg 1,3 –score-min L,-0.6,-1 -a). However, even using these custom parameters, bowtie2 fails to detect alignments of ≈20% of Illumina reads, resulting in an underestimation of the heights of long reads. To compensate for this effect, we applied an adjustment factor $\frac{100}{100-20} = 1.25$ to artificially inflate the heights in our formula for estimating the metagenome capacity.

Currently, there is a shortage of publicly available hybrid metagenomics datasets (containing both Illumina and SMS reads). Ideally, Illumina and SMS reads for such datasets should be generated at the same time so that the abundances of individual genomes within a metagenome are the same for Illumina and SMS reads, implying that the depth of coverage by Illumina reads is proportional to the depth of coverage by SMS reads. In practice, since the SMS reads for these datasets were often generated as an afterthought, Illumina and SMS reads for the publicly available hybrid metagenomics datasets are generated at different time points and prepared for sequencing using different sample preparation protocols. Thus, since metagenome composition is changing and is subject to blooms [33], the existing hybrid datasets do not necessarily feature the proportional depths of coverage by Illumina and SMS reads. Our analysis revealed that the fractions of Illumina and SMS reads aligned to each of the reference genomes for publicly available hybrid synthetic metagenomic dataset may differ by two orders of magnitude. This difference in the genome coverages by short and long reads in the publicly available hybrid metagenomics datasets makes our approach inapplicable to the currently available hybrid metagenomics datasets.

References

1. Amann, R., Rosselló-Móra, R.: After all, only millions? mBio **7**(4), e00,99916 (2016)
2. Bankevich, A., Nurk, S., Antipov, D., et al.: SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**(5), 455–477 (2012)
3. Bankevich, A., Pevzner, P.A.: TruSPAdes: barcode assembly of TruSeq synthetic long reads. Nat. Methods **13**, 248–250 (2016). https://doi.org/10.1038/nmeth.3737
4. Capo, E., Debroas, D., Arnaud, F., Domaizon, I.: Is planktonic diversity well recorded in sedimentary DNA? Toward the reconstruction of past protistan diversity. Microb. Ecol. **70**(4), 865–875 (2015)
5. Chao, A., Bunge, J.: Estimating the number of species in a stochastic abundance model. Biometrics **58**(3), 531–539 (2002). https://doi.org/10.1111/j.0006-341X.2002.00531.x
6. Chen, Y., Kuang, J., Jia, P., Cadotte, M.W., Huang, L., Li, J., Liao, B., Wang, P., Shu, W.: Effect of environmental variation on estimating the bacterial species richness. Front. Microbiol. **8**, 690 (2017)
7. Compeau, P.E.C., Pevzner, P.A., Tesler, G.: How to apply de Bruijn graphs to genome assembly. Nat. Biotechnol. **29**(11), 987–991 (2011). https://doi.org/10.1038/nbt.2023

8. Curtis, T.P., Sloan, W.T., Scannell, J.W.: Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. U.S.A. **99**(16), 10494–10499 (2002). https://doi.org/10.1073/pnas.142680199
9. Driscoll, C.B., Otten, T.G., Brown, N.M., Dreher, T.W.: Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. Stand. Genomic Sci. **12**(1), 9 (2017)
10. Dykhuizen, D.E.: Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek **73**(1), 25–33 (1998)
11. Ellegaard, K.M., Engel, P.: Beyond 16S rRNA community profiling: intra-species diversity in the gut microbiota. Front. Microbiol. **7**, 1475 (2016)
12. Frisli, T., Haverkamp, T.H.A., Jakobsen, K.S., Stenseth, N.C., Rudi, K.: Estimation of metagenome size and structure in an experimental soil microbiota from low coverage next-generation sequence data. J. Appl. Microbiol. **114**(1), 141–151 (2013). https://doi.org/10.1111/jam.12035
13. Gao, W., Weng, J., Gao, Y., Chen, X.: Comparison of the vaginal microbiota diversity of women with and without human papillomavirus infection: a cross-sectional study. BMC Infect. Dis. **13**(1), 271 (2013)
14. Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G.: QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**(8), 1072–1075 (2013). https://doi.org/10.1093/bioinformatics/btt086
15. Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., Weitz, J.S.: Robust estimation of microbial diversity in theory and in practice. ISME J. **7**(6), 1092–1101 (2013). https://doi.org/10.1038/ismej.2013.10
16. Haider, B., Ahn, T.H., Bushnell, B., et al.: Omega: an overlap-graph de novo assembler for metagenomics. Bioinformatics **30**(19), 2717–2722 (2014). https://doi.org/10.1093/bioinformatics/btu395
17. Hong, S.H., Bunge, J., Jeon, S.O., Epstein, S.S.: Predicting microbial species richness. Proc. Natl. Acad. Sci. U.S.A. **103**(1), 117–122 (2006). https://doi.org/10.1073/pnas.0507245102
18. Hooper, S.D., Dalevi, D., Pati, A., Mavromatis, K., Ivanova, N.N., Kyrpides, N.C.: Estimating DNA coverage and abundance in metagenomes using a gamma approximation. Bioinformatics **26**(3), 295–301 (2010). https://doi.org/10.1093/bioinformatics/btp687
19. Hughes, J.B., Hellmann, J.J., Ricketts, T.H., Bohannan, B.J.: Counting the uncountable: statistical approaches to estimating microbial diversity. Appl. Environ. Microbiol. **67**(10), 4399–4406 (2001)
20. Jousset, A., Bienhold, C., Chatzinotas, A., et al.: Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J. **33**(4), 853–862 (2017)
21. Kashtan, N., Roggensack, S.E., Rodrigue, S., et al.: Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus. Science **344**(6182), 416–420 (2014)
22. Kemp, P.F., Aller, J.Y.: Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. FEMS Microbiol. Ecol. **47**(2), 161–177 (2004). https://doi.org/10.1016/S0168-6496(03)00257-5
23. Kuleshov, V., Jiang, C., Zhou, W., Jahanbani, F., Batzoglou, S., Snyder, M.: Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat. Biotechnol. **34**(1), 64–69 (2015). https://doi.org/10.1038/nbt.3416
24. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–359 (2012)

25. Lennon, J.T., Locey, K.J.: The underestimation of global microbial diversity. mBio **7**(5), e01,298-16 (2016). https://doi.org/10.1128/mBio.01298-16
26. Lennon, J.T., Placella, S.A., Muscarella, M.E.: Relic DNA contributes minimally to estimates of microbial diversity. bioRxiv, p. 131284 (2017)
27. Li, R., Hsieh, C.L., Young, A., et al.: Illumina synthetic long read sequencing allows recovery of missing sequences even in the "finished" C. elegans genome. Sci. Rep. **5**, 10,814 (2015). https://doi.org/10.1038/srep10814
28. Lladser, M.E., Gouet, R., Reeder, J.: Extrapolation of urn models via poissonization: accurate measurements of the microbial unknown. PLoS ONE **6**(6), e21,105 (2011). https://doi.org/10.1371/journal.pone.0021105
29. Locey, K.J., Lennon, J.T.: Scaling laws predict global microbial diversity. Natl. Acad. Sci. U.S.A. **113**(21), 5970–5975 (2016)
30. Loose, M., Malla, S., Stout, M.: Real-time selective sequencing using nanopore technology. Nat. Methods **13**(9), 751–754 (2016)
31. Lynch, M.D.J., Neufeld, J.D.: Ecology and exploration of the rare biosphere. Nat. Rev. Microbiol. **13**(4), 217–229 (2015). https://doi.org/10.1038/nrmicro3400
32. McCoy, R.C., Taylor, R.W., Blauwkamp, T.A., et al.: Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoS ONE **9**(9), e106,689 (2014). https://doi.org/10.1371/journal.pone.0106689
33. McDonald, D., et al.: American gut: an open platform for citizen-science microbiome research (2018, submitted)
34. Miller, C.S., Baker, B.J., Thomas, B.C., Singer, S.W., Banfield, J.F.: Emirge: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. **12**(5), R44 (2011). https://doi.org/10.1186/gb-2011-12-5-r44
35. Pedrós-Alió, C., Manrubia, S.: The vast unknown microbial biosphere. Proc. Natl. Acad. Sci. U.S A. **113**(24), 6585–6587 (2016). https://doi.org/10.1073/pnas.1606105113
36. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.: A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464**(7285), 59–65 (2010)
37. Rodriguez-R, L.M., Konstantinidis, K.T.: Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics **30**(5), 629–635 (2014). https://doi.org/10.1093/bioinformatics/btt584
38. Roesch, L.F.W., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K.M., Kent, A.D., Daroub, S.H., Camargo, F.A.O., Farmerie, W.G., Triplett, E.W.: Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1**(4), 283–290 (2007). https://doi.org/10.1038/ismej.2007.53
39. Rosselli, R., Romoli, O., Vitulo, N., et al.: Direct 16S rRNA-SEQ from bacterial communities: a PCR-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon. Sci. Rep. **6**, 32,165 (2016)
40. Rozov, R., Brown Kav, A., Bogumil, D., Shterzer, N., Halperin, E., Mizrahi, I., Shamir, R.: Recycler: an algorithm for detecting plasmids from de novo assembly graphs. Bioinformatics **33**(4), 475–482 (2017)
41. Scher, J.U., Ubeda, C., Artacho, A., et al.: Decreased bacterial diversity characterizes the altered gut microbiota in patients with psoriatic arthritis, resembling dysbiosis in inflammatory bowel disease. Arthritis Rheumatol. **67**(1), 128–139 (2015). https://doi.org/10.1002/art.38892

42. Schloss, P.D., Girard, R.A., Martin, T., Edwards, J., Thrash, J.C.: Status of the archaeal and bacterial census: an update. mBio **7**(3), e00,201-16 (2016). https://doi.org/10.1128/mBio.00201-16
43. Schloss, P.D., Handelsman, J.: Status of the microbial census. Microbiol. Mol. Biol. Rev. **68**(4), 686–691 (2004). https://doi.org/10.1128/MMBR.68.4.686-691.2004
44. Shade, A.: Diversity is the question, not the answer. ISME J. **11**(1), 1–6 (2016). https://doi.org/10.1038/ismej.2016.118
45. Shakya, M., Quince, C., Campbell, J.H., Yang, Z.K., Schadt, C.W., Podar, M.: Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. Environ. Microbiol. **15**(6), 1882–1899 (2013). https://doi.org/10.1111/1462-2920.12086
46. Sharon, I., Kertesz, M., Hug, L.A., et al.: Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. **25**(4), 534–543 (2015). https://doi.org/10.1101/gr.183012.114
47. Sharpton, T.J., Riesenfeld, S.J., Kembel, S.W., et al.: PhyLOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PLoS Comput. Biol. **7**(1), e1001,061 (2011)
48. Sogin, M.L., Morrison, H.G., Huber, J.A., et al.: Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U.S.A. **103**(32), 12115–12120 (2006). https://doi.org/10.1073/pnas.0605127103
49. Sunagawa, S., DeSantis, T.Z., Piceno, Y.M., et al.: Bacterial diversity and White Plague Disease-associated community changes in the Caribbean coral Montastraea faveolata. ISME J. **3**(5), 512–521 (2009). https://doi.org/10.1038/ismej.2008.131
50. Taur, Y., Jenq, R.R., Perales, M.A., et al.: The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. Blood **124**, 1174–1182 (2014). https://doi.org/10.1182/blood-2014-02-554725
51. Tiedje, J.: Microbial diversity: of value to whom? ASM News **60**, 524–525 (1994)
52. Voskoboynik, A., Neff, N.F., Sahoo, D., et al.: The genome sequence of the colonial chordate, Botryllus schlosseri. eLife **2**, 69 (2013). https://doi.org/10.7554/eLife.00569
53. White, R.A., Bottos, E.M., Roy Chowdhury, T., et al.: Moleculo long-read sequencing facilitates assembly and genomic binning from complex soil metagenomes. mSystems **1**(3) (2016). https://doi.org/10.1128/mSystems.00045-16
54. Williamson, M., Gaston, K.J.: The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. J. Anim. Ecol. **74**(3), 409–422 (2005). https://doi.org/10.1111/j.1365-2656.2005.00936.x
55. Willis, A.: Extrapolating abundance curves has no predictive power for estimating microbial biodiversity. Proc. Natl. Acad. Sci. U.S.A. **113**(35), E5096 (2016). https://doi.org/10.1073/pnas.1608281113