International Journal of Food Microbiology 410 (2024) 110488

Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

Review

A user's guide to the bioinformatic analysis of shotgun metagenomic sequence data for bacterial pathogen detection

Blake G. Lindner [a,1], Kenji Gerhardt [b,1], Dorian J. Feistel [b,1], Luis M. Rodriguez-R [c], Janet K. Hatt [a], Konstantinos T. Konstantinidis [a,b,*]

[a] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[b] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
[c] Department of Microbiology, Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Austria

ARTICLE INFO

Keywords:
Limit of detection
Relative abundance
Metagenome
Read recruitment plots
Average nucleotide identity
Bioinformatics
Foodborne pathogens

ABSTRACT

Metagenomics, i.e., shotgun sequencing of the total microbial community DNA from a sample, has become a mature technique but its application to pathogen detection in clinical, environmental, and food samples is far from common or standardized. In this review, we summarize ongoing developments in metagenomic sequence analysis that facilitate its wider application to pathogen detection. We examine theoretical frameworks for estimating the limit of detection for a particular level of sequencing effort, current approaches for achieving species and strain analytical resolution, and discuss some relevant modern tools for these tasks. While these recent advances are significant and establish metagenomics as a powerful tool to provide insights not easily attained by culture-based approaches, metagenomics is unlikely to emerge as a widespread, routine monitoring tool in the near future due to its inherently high detection limits, cost, and inability to easily distinguish between viable and non-viable cells. Instead, metagenomics seems best poised for applications involving special circumstances otherwise challenging for culture-based and molecular (e.g., PCR-based) approaches such as the de novo detection of novel pathogens, cases of co-infection by more than one pathogen, and situations where it is important to assess the genomic composition of the pathogenic population(s) and/or its impact on the indigenous microbiome.

1. Key parameters in shotgun metagenomic datasets

Several key concepts and assumptions that are central to all metagenomic approaches must be clearly defined to avoid confusion. First, metagenomics is an approach that involves the application of shotgun sequencing to the total DNA of a microbial community. Producing data in this way yields a set of sequence reads thought to represent the features (e.g., gene, operon, pathway, or genome) at the same proportions relative to one another in which they existed at time of sampling (Handelsman et al., 2007). This assumption can be violated due to inadequate or biased sampling, DNA extraction biases, and G + C% content bias in protocols that require amplification prior to sequencing, to name only a few of the potential limiting factors. Several studies have shown that biases can be managed and efforts to do so are certainly the responsibility of all researchers utilizing metagenomics (Jones et al., 2015; Nearing et al., 2021). These issues have been covered extensively elsewhere and are not discussed further here (McLaren et al., 2019; Nayfach and Pollard, 2016). Instead, the focus of this review is the bioinformatic analysis of the resulting sequence data based on the assumption that the sequence data is adequately representative of the sampled communities.

Three central metrics associated with metagenomic datasets are sequencing depth, sequencing breadth, and sequencing effort (Table 1). Sequencing depth refers to the number of metagenomic reads covering a feature at a base pair position and is often communicated as an average for the entire feature followed by an "X" (e.g., 10× depth). Sequencing depth may sometimes be referred to as simply "coverage" (e.g., 10× coverage). In contrast, sequencing breadth denotes the fraction of base pair positions in a feature covered by at least one metagenomic read and is often communicated as a fraction or percentage (e.g., 50 % breadth). Sequencing breadth is unfortunately also sometimes referred to as "coverage" in various studies, which can lead to confusion (e.g., 50 %

* Corresponding author at: School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA.
E-mail address: kostas@ce.gatech.edu (K.T. Konstantinidis).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijfoodmicro.2023.110488
Received 17 May 2023; Received in revised form 15 September 2023; Accepted 11 November 2023
Available online 17 November 2023

Table 1
Key terms and their use in estimating metagenomic sensitivity.

Sequencing depth	The number of sequence reads covering a feature at a particular base pair position or the average across all base pair positions of a feature (e.g., 10×)
Sequencing breadth	The fraction of base pair positions in a feature covered by at least one sequence read (e.g., 50 %, 0.5).
Sequencing effort	The amount of data produced by metagenomic sequencing per sample (e.g., 5 gigabases (or Gbp), 33 million reads).
Genome equivalents	The average sequencing depth of universal single copy gene(s) in a sample (e.g., 1000 GEQ).
Relative abundance	The ratio of some metric representing a feature's average sequencing depth to an appropriate metric for the sample's total sequencing effort (e.g., 1 %, 0.01).

(1) $S_{min} = \rho_{LOD}\frac{1}{\alpha_{target}}\gamma_{target}$ (2) $S_{min} = \rho_{LOD}\frac{C_{total}}{C_{target}}\gamma_{avg}$ (3) $C_{min} = \frac{\rho_{LOD}}{GEQ}C_{total}$

Top: Definitions of key terms related to estimating metagenome sensitivity. Bottom: Several expressions for estimating relationships between sequencing depth and population detection in metagenomic datasets. S_{min} (in base pairs) defines the minimum expected sequencing effort to detect a target genome. ρ_{LOD} is the sequencing depth of the target genome considered necessary to call a true detection. α_{target} is the sequence-based relative abundance of the target genome in situ. C_{target}/C_{total} is the cell-based relative abundance of the target genome in situ with C_{target} and C_{total} being the target genome and total community cellular concentrations, respectively. γ represents the genome size (in base pairs) of the target genome or the average genome size of the community. Lastly, C_{min} is the minimum cellular load expected to be detectable given by choice of ρ_{LOD} and genome equivalents (GEQ).

coverage). Herein, we only use the explicit terms sequencing depth and sequencing breadth to avoid ambiguity. The amount of data produced by metagenomic sequencing is often referred to as sequencing effort and is usually measured on the order of gigabases (i.e., 10^9 base pairs or Gbp) per sample. Sequencing effort is commonly communicated as the number of total reads or base pairs in a dataset. Alternatively, sequencing effort can be indirectly communicated in terms of the sequence depth of a universal single copy gene (e.g., RNA polymerase subunit B or *rpoB*) or the average depths of a set of universal prokaryotic single copy genes. This approach for reporting sequencing effort is often termed "genome equivalents" (GEQ) and represents the equivalent number of prokaryotic genomes present in a metagenome. This metric is useful because it pairs the idea of sequencing effort with the average genome size of the prokaryotic community sampled.

Ratios of sequencing depth and sequencing effort produce one of the more useful metrics reported in metagenomics: relative abundance. Although multiple methods are commonly used to estimate the relative abundances of a feature, it is important to note that it is always produced via some ratio of a feature's sequencing depth to the sample's sequencing effort. Accordingly, the sum of the relative abundances of all features in a dataset equals to one, and methods for calculating these abundances are discussed in subsequent sections. Importantly, relative abundances are the proportions representing various features in metagenomic datasets and can hold powerful clues about the underlying biology and ecology of the community being studied. It is important to note, however, that relative abundances are proportions and as such cannot indicate the actual load of a feature within a sample without additional data or information about the system (Morton et al., 2019). Therefore, considering the inherent compositionality of metagenomic datasets as well as the notion that any given read set represents only a subset of the total community DNA are both important outlooks to maintain when researchers aim to use metagenomics to detect pathogens. The next section summarizes approaches and recent findings that can help address some of the abovementioned challenging aspects related to metagenomic work.

2. Prokaryotic species may exist and be recognized by metagenomic data

Many studies have used species as the main target unit for organizing and interpreting metagenomic data but to what extent species exist for prokaryotes and can be recognized by metagenomic data remains a highly debatable issue. Recent evidence, however, has lent strong support for using species as a key unit of diversity. Specifically, metagenomic studies of natural bacterial populations (Konstantinidis and DeLong, 2008; Olm et al., 2020) and comparative analysis of the genomes of bacterial isolates (Jain et al., 2018; Rodriguez et al., 2021) have revealed that bacteria and archaea predominantly form sequence discrete species. Specifically, these studies have shown that intra-species genomic sequence relatedness typically ranges from 95 % to 100 % genome-aggregate average nucleotide identity (ANI) depending on the population considered. For clarity, ANI refers to the average nucleotide identity of all shared genes between the two genomes compared. These results mean that since the last diversity sweep event, younger species show lower levels of intra-species diversity compared to older species. In contrast, ANI values between members of distinct bacterial species are typically lower than 90 % (Caro-Quintero and Konstantinidis, 2012).

These results contrast with the view that bacteria do not form distinct species due to the extensive genetic exchange they often undergo and their very large population sizes (e.g., species diversity sweep events are not common) (Doolittle, 2019; Lawrence, 2002). The difference between the metagenomics results and those of previous studies that reported non-discrete species, including for important foodborne pathogens such as *Escherichia coli* (Luo et al., 2011) and *Campylobacter jejuni* (Sheppard et al., 2008), may be due to isolation biases (Caro-Quintero and Konstantinidis, 2012; Rodriguez et al., 2021). That is, the latter studies included heterogeneous collections of organisms isolated in the laboratory that represent different ecological niches and genomic adaptations specific to local environmental conditions at the place of isolation. Comparisons among such organisms may have confounded the existence of sequence discrete species by often revealing non-discrete species (Hanage et al., 2005; Luo et al., 2011). However, a more recent analysis of all available isolate genomes in the NCBI database (n = ~90,000) also revealed sequence discontinuities between most named species around 85–95 % ANI (Jain et al., 2018), consistent with the picture emerging from metagenomics but not the early picture from comparisons of isolate genomes that often showed indiscrete species.

In summary, recent metagenomic and genomic high-throughput studies have revealed that bacterial species may exist, which is an important prerequisite for species and strain detection – especially within the context of culture independent methodologies like metagenomics. Further, how genomes have been classified into named species during recent decades is highly consistent with the ANI threshold; ~97 % of named species encompass only genomes sharing >95 % ANI. Notably, similar results to those mentioned above for bacteria have recently been reported for other microbes, most notably, protozoa (Seabolt et al., 2021) and viruses (Simmonds et al., 2017), indicating a broad applicability of the 95 % ANI threshold within the microbial world.

3. Reliable units of intra-species diversity based on ANI

Despite the likely existence of sequence discrete species among prokaryotes, foodborne outbreaks are typically caused by specific members (e.g., strains) of a species, and thus it is important to obtain intra-species resolution in cases where metagenomic approaches are employed for pathogen detection and to distinguish pathogens from any innocuous relatives of the same or closely related species co-occurring in

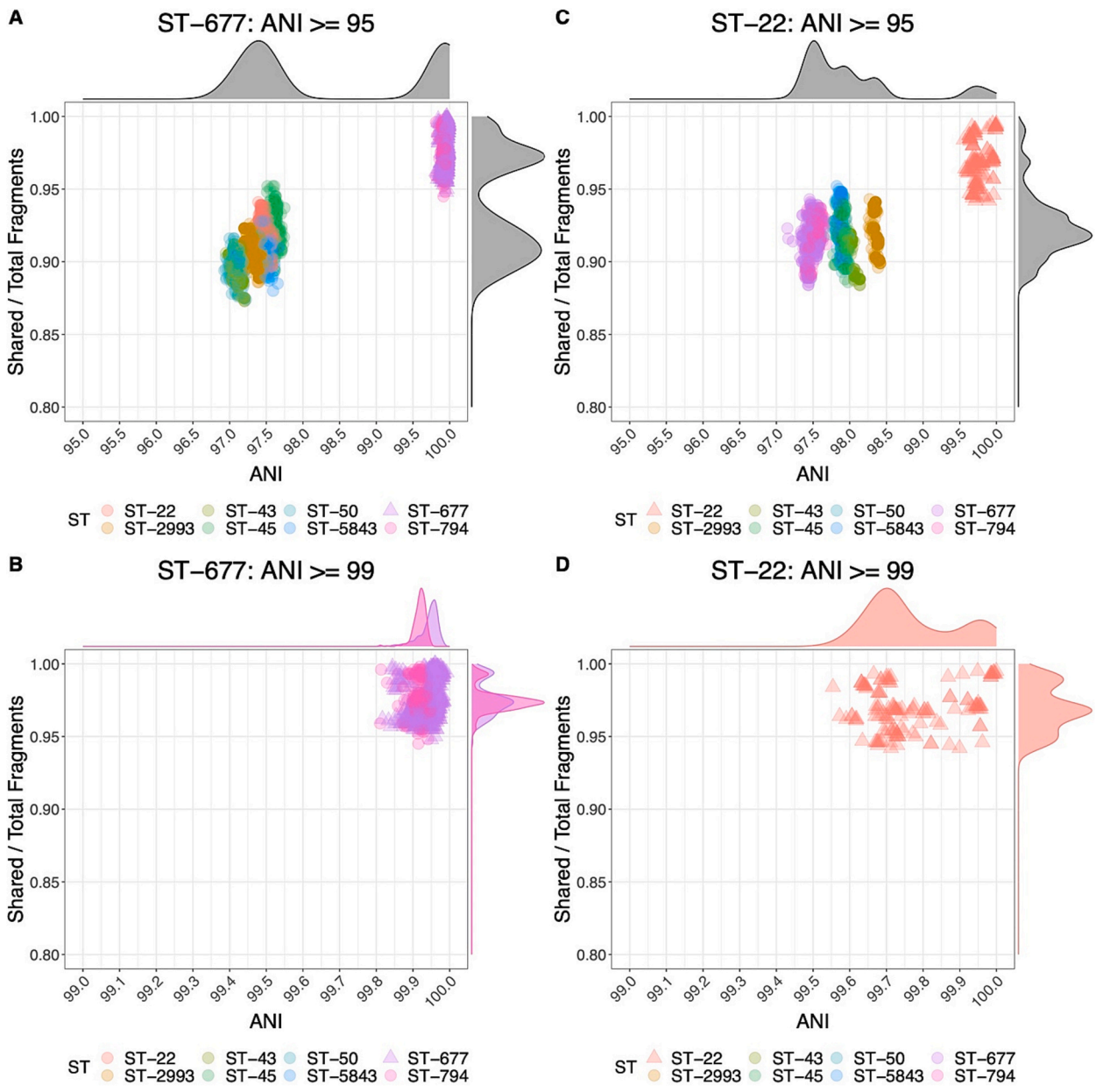
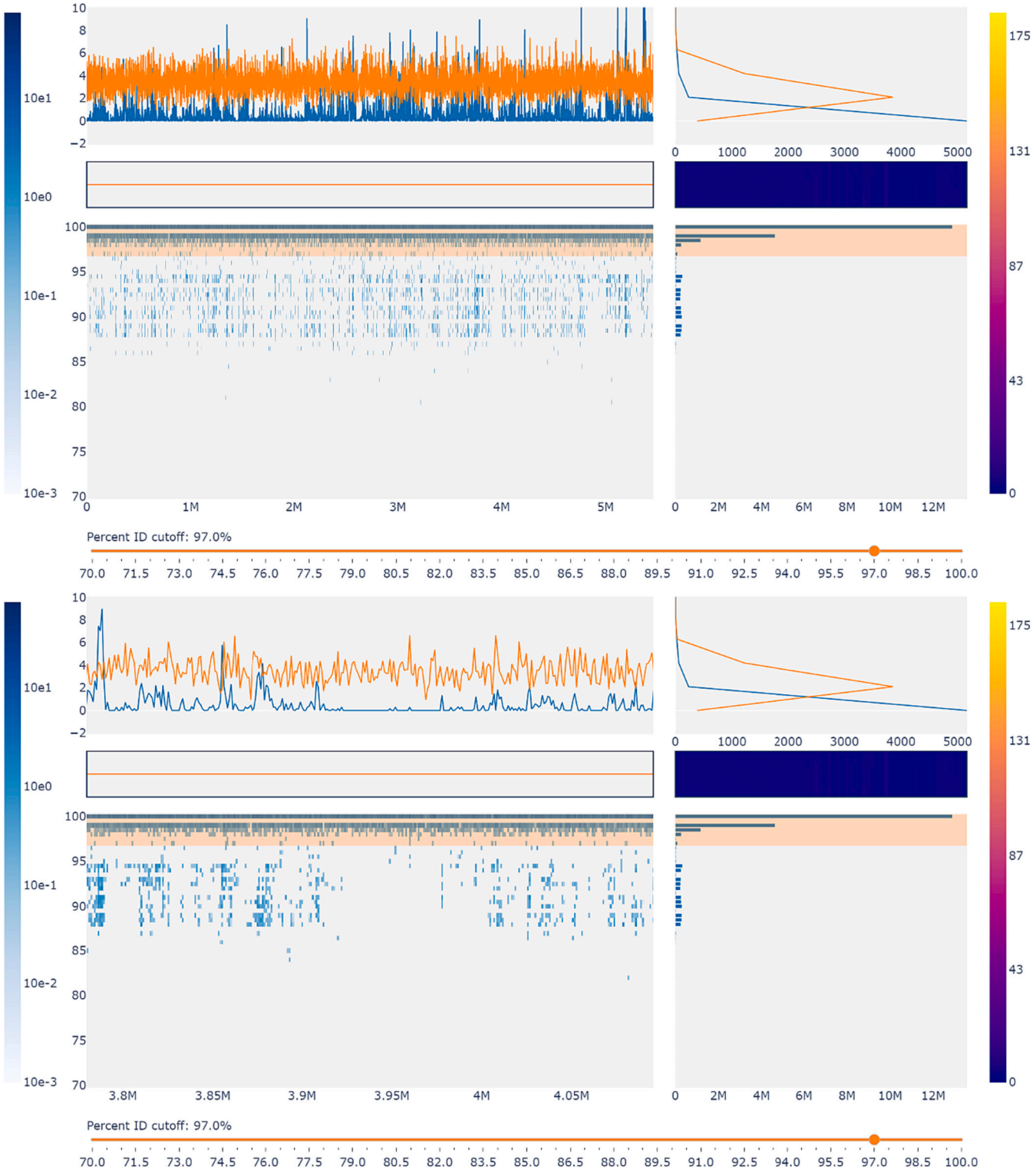


Fig. 1. Comparison of the ANI threshold with Sequence Types (STs). The graph is based on 239 complete *Campylobacter jejuni* genomes downloaded from NCBI and shows their pairwise shared genome fraction (y-axes) plotted against their ANI values (x-axes). The top eight STs consisting of the greatest number of genomes were used for this analysis (ST-677 = 42 genomes, ST-22 = 17, ST-2993 = 14, ST794 = 12, ST-43 = 11, ST-21 = 10, ST-50 = 9, and ST-45 = 8). FastANI was used to generate pairwise ANI values and shared genome fraction with default settings, excluding self-comparisons. The genome pairs compared are colored based on the ST that the two genomes in a pair were assigned to, using the tool mlst (https://github.com/tseemann/mlst). The filtered data was then plotted in R-studio using the *ggplots* package with either ST-677 (Panels A and B) or ST-22 (Panels C and D) as the reference ST, for which the reference for a given plot is indicated by a triangle in its respective legend. Panels B and D represent zoomed-in versions of Panel A and C. Marginal density plots show either the joint (grey) or individual (color) distribution for STs. Note that the 99.5 % ANI threshold closely matches the areas of discontinuity between distinct STs (e.g., Panel A and C) and that probably STs 667 and 794 are too overlapping and clonal in terms of the genomic relatedness of the grouped genomes, and thus could be merged into one ST (Panel B).

the sample. An important associated challenge, however, is the precise definition of "strain" particularly within the context of metagenomic datasets possibly lacking associated cultured-based data. Prokaryotic taxonomy defines strains on the basis of the pure culture approach as "*a group of genetically similar descendants of a single colony or cell*" (Parker et al., 2019). Therefore, a strain includes all derivative lines of a single isolate, even if these descendants have undergone substantial gene loss, gene duplications, or genomic rearrangements. Such mutational events are common when an organism is grown in the laboratory and commonly result in significant phenotypic differences (Knoppel et al., 2018); yet, despite clear structural and even functional dissimilarities, the wild-type and the lab-adapted cells are usually considered the same strain (Dijkshoorn et al., 2000). Although, if the gene differences involve the key phenotype of interest such as virulence or antibiotic resistance, the derived culture could be designated as different strain. Perhaps more importantly, in surveys of natural populations, where strain ancestry information is typically unavailable, strains have instead been discerned based on single nucleotide variants patterns (SNVs). However, a widely accepted definition on the number of SNVs expected to define a strain has not yet emerged, even for such natural populations (Yan et al., 2020).

Recently Rodriguez-R and colleagues analyzed 330 bacterial species



(caption on next page)

Fig. 2. An example of the new read-recruitment plot. For these plots, a healthy human gut metagenome (accession SRX20110658) was spiked with simulated reads from a pathogenic *E. coli* genome (accession GCA_026384595.1), and resulting reads were mapped against the latter genome (reference). The top plot shows the whole genome; bottom plot shows a zoomed-in version around positions 3.9Mbp to 4.0Mbp, where a pathogenicity-associated genomic island is located. The four different panels of each plot represent: Bottom-Left: is a 2-D histogram displaying the percent identity of reads mapped to the reference genome on the y-axis and the position in the genome on the x-axis. Shading in orange represents reads mapping above the nucleotide identity threshold, here shown at 97 % (tool's default is 95 %). Top-Left: is a line plot of the average sequencing depth across the genome region shown on the main panel (bottom-left). The orange line represents reads mapping above the nucleotide identity threshold; blue represents reads below the threshold. Top-right: is a histogram of sequencing depths across the reference sequence, with colors corresponding to those in the top-left panel. Bottom-right: is a histogram of the number of bases displayed in the main panel that fall within specific percent identity windows (y-axis), displayed in log scale. Minor panels in the middle: the middle two subplots are for gene annotations (left), and per-contig sequencing depth (TAD) summaries (right). The annotation plot will include information about the genes predicted in the genome when available (off by default), including strand, G + C % content, and coordinates, while TAD80 values (and other sequencing depth metrics) are provided for the contigs. Please see tool's manual and interactive mode (pop-up boxes) for further details.
Note the even sequencing depth over the entire genome caused by the spike-in reads of the pathogenic *E. coli* (top-left panel) and high ANIr (~100 %, bottom-right panel) compared to the more variable sequencing depth and lower ANIr (~88–93 %) from reads representing the close relative(s) present in the metagenome in orange. Also note the regions of near-zero sequencing depth (i.e., corresponding to genes/regions that are not detectable) for the genomic island near the bottom of the top-left panel for the population of the relative(s) (but not the spiked-in pathogen). See also the text for further details.

that are each well-sampled by multiple sequenced isolate genomes, representing both pathogenic and environmental taxa, and showed that a clear bimodal distribution in the ANI values characterizes most of these species (or 95 % ANI-defined genomospecies). That is, there is a scarcity of genome pairs (or an ANI gap) showing 99.2–99.8 % ANI (midpoint at 99.5 % ANI) in contrast to genome pairs showing ANI >99.8 % or < 99.2 % (Fig. 1) (Rodriguez-R et al., In press). Therefore, it appears that another important level of genomic differentiation may exist within species and can be used to define and standardize intra-species units across different taxa. Further, the 99.5 % ANI gap is highly consistent with sequence types (STs), a key concept that has been widely used, especially in medical microbiology and epidemiological studies, to identify an outbreak caused by a specific pathogenic organism (strain) or groups of highly related organisms. An ST is typically defined as a group of genomes with no nucleotide sequence differences in 6–7 selected genetic loci (Maiden et al., 1998). Their evaluation based on species that are well sampled by genome sequences such as *Escherichia coli* and *Campylobacter jejuni* has shown that the 99.5 % ANI threshold is largely consistent with how STs have been defined in these species, e.g., ~80 % of ST assignments were supported by the 99.5 % ANI threshold. Nonetheless, the 99.5 % ANI threshold provides intra-species groups with ~20 % higher accuracy in terms of genomic and gene-content relatedness of the grouped genomes (Fig. 2 and Rodriguez-R et al., In press). The main reason for the higher accuracy of the 99.5 % ANI threshold is the limited signal carried by 6–7 loci vs. the whole genome (ANI), including the horizontal transfer of one (or more) of the loci that can confound ST assignments. Other important advantages of the 99.5 % ANI approach are that it can be automatically implemented, and thus does not require manual curation, which is needed for establishing ST numbers for new sequences not seen previously (Maiden et al., 1998). Additionally, the computation of ANI is two orders of magnitude faster compared to the phylogenetic placement of a genome using all core (but probably not faster than using 6-7 marker) genes in whole-genome-based ST analysis (Jain et al., 2018).

Rodriguez-R and colleagues did not observe another pronounced ANI gap within species, and thus recommended the use of the term strain only for nearly identical genomes. Specifically, they proposed to define a strain as a collection of genomes sharing ANI >99.99 % based on the high gene-content similarity observed at this level based on the genomes compared e.g., typically, >99.0 % gene content is shared (Viver et al., 2023). It should be noted, however, that genomes sharing >99.99 % ANI are not clonal, meaning they may still show non-trivial gene-content differences attributed (mostly) to mobile elements (e.g., 1 % gene-content difference for a 5Mbp genome translates to 50 genes being different). It was suggested to let the ANI >99.99 % threshold override such mobile-element-driven gene-content differences in order to simplify strain identification and communication. However, in cases where important phenotypic differences that distinguish between organisms sharing ANI >99.99 % are known such as antibiotic resistance genes carried by plasmids, the proposed definition for strain could be neglected or adjusted upwards as appropriate. Accordingly, Rodriguez-R and colleagues proposed to use the 99.5 % ANI threshold to define new or refine existing STs toward more genomically homogenous and data-driven STs. If the ST should maintain its original conception of 6–7 identical loci for historic or other reasons, they suggested instead to use the term genomovar to refer to these 99.5 %-ANI intra-species units. The term genomovar was originally used to name distinct genomic groups within species that cannot be distinguished phenotypically from each other in order to be named as distinct species (Ursing et al., 1995). Hence, genomovar may best capture conceptually the 99.5 % ANI units.

Notably, discrete, or somewhat discrete (Hanage et al., 2005), ecological or evolutionary units within bacterial species have long been recognized and are designated by various terms such as ecotypes, clonal complexes, sequence types, and serotypes, among several other terms (recently reviewed in Rossello-Mora and Amann, 2015; Van Rossum et al., 2020). However, the application of these units has commonly been inconsistent between different taxa and studies, e.g., different marker genes and standards for each marker are used, creating challenges in communication about intra-species diversity. The findings by Rodriguez-R and colleagues (e.g., Fig. 1 and Rodriguez-R et al., In press) suggest that the 99.5 % ANI clusters could represent such a consistent intra-species unit that can be used to help standardize definitions across taxa. Therefore, these ANI thresholds can provide convenient and robust means in identifying strains and STs/genomovars associated with food poisoning and outbreaks, as well as facilitate communication about these intra-species units. While the 99.2–99.8 % ANI range should represent the gap for most species based on the dataset evaluated (all 330 species evaluated are available on the GitHub at https://github.com/rotheconrad/bacterial_strain_definition), it was still suggested to directly evaluate the ANI value distribution among genomes of the species of interest and adjusting the genomovar/ST-defining ANI threshold to match the gap in the observed ANI value distribution should the data indicate that a 99.5 % cutoff is inappropriate.

Detecting STs and strains based on short-read sequencing data (reads of 100-250bp in length) and read-recruitment plots is technically challenging because such short-reads do not provide enough resolution in the most critical area, that of 99–100 % nucleotide identity level (e.g., 1 mismatch in a 200bp read would result in a nucleotide identity of 99.5 %, lacking resolution in the critical 99.5–100 % nucleotide identity range). However, long-read sequencing such as that offered by the Oxford Nanopore and PacBio instruments (reads 10Kbp or longer), or isolate and single-cell complete or draft genomes, should provide adequate resolution as our recent work has shown (Rodriguez-R et al., In press). The approach outlined above should provide the means to identify distinct units within species and define strains and other sub-species units. Further, several tools such as inStrain (Olm et al., 2021), ConStrains (Luo et al., 2015), and StrainGE (van Dijk et al., 2022) to name a few, can reliably identify strains present in a short-read

metagenome based on the nucleotide substitution patterns when the corresponding strain genomes are available in the internal databases of these tools. However, these tools are limited when novel strains are instead present in the metagenome, and cannot typically reconstruct the genomes of (known) strains in cases where multiple strains or strain hybrids are present (which is often the case) due to the nature of short-read data. Additionally, these tools have varied requirements regarding the minimum sequencing depth necessary to accurately identify strains to which users should pay strict attention. Finally, it should also be mentioned that the exact mechanisms underlying the 99.5 % ANI gap, or the previously established 95 % ANI for the species level, remain unclear but are likely related to ecological differentiation, coupled to recombination frequency, e.g., higher horizontal gene transfer mediated by homologous recombination within vs. between groups (Fraser et al., 2009; Hanage et al., 2005), and should be the subject of future research.

In summary, the ANI-based definitions proposed by Rodriguez-R and colleagues for species (>95 % ANI), genomovar (>99.5 %), and strain (>99.99 %) designations should provide convenient and reproducible means for grouping sequence reads into clear units. Additionally, these definitions facilitate communication about diversity units during pathogen monitoring and/or outbreak detection. Next, we will explore technical approaches for utilizing these concepts and their associated ANI thresholds for pathogen detection and diagnosis.

4. Detecting sequence discrete units (species and strains) in metagenomic datasets

Several different approaches exist for monitoring species and strains within metagenomic datasets. A core feature shared among all approaches is the need for some kind of reference sequence (e.g., a diagnostic gene or genome) to which sequence reads can be compared in order to identify the corresponding taxon. These reference sequences can be obtained from existing data (e.g., Benson et al., 2012), from sequencing efforts of isolates or enrichments, or produced de novo through recovery of metagenome-assembled genomes (MAGs) from the metagenomic data itself (Fig. 4). MAGs are typically thought to represent the average, composite genome of a population present in a sample (Chen et al., 2020; Sczyrba et al., 2017). The choice of what reference sequence should be used is undoubtedly a key parameter that depends, at least, on both the pathogen(s) being surveyed and the background sample matrix. Yet, the tendency of prokaryotes to form sequence discrete units when considering whole genome ANI relatedness is incredibly useful for evaluating what signals in metagenomic data likely correspond to populations of interest.

Taxonomic profilers are one of the most accessible approaches for surveying the prevalence (and abundance) of microbial species in metagenomic datasets. These tools are invariably based on some notion of sequence discreteness occurring in the underlying data, whether using percent sequence identity cut-offs or k-mer matching, to determine what reads match to what taxa in the reference dataset. Performance of any taxonomic profiler primarily relies on the reference database construction and searching tool/approach, in addition to a few parameters that are further discussed in the next section. Popular examples of metagenomic profiling software include MetaPhlAn (Beghini et al., 2021; Blanco-Miguez et al., 2023), Kaiju (Menzel et al., 2016), and Kraken (Lu and Salzberg, 2020), and these types of software are usually heterogeneous with respect to how they handle tied matches among sequence reads, database construction, and benchmarking efforts. To assess the importance of these differences on the results obtained, some tools have been benchmarked across a few use cases. It is important to note, however, that these cases may or may not correspond with the performance expected in food systems, which should be a key consideration of researchers aiming to use taxonomic profilers in food sample matrices. Notably, community efforts are ongoing which aim to more comprehensively understand and benchmark taxonomic profiler performance across a variety of representative use cases (Sczyrba et al., 2017). Importantly, taxonomic profilers report relative abundances for detected populations though these methods are not uniform between approaches and a tool's sensitivity is usually not clear to the user, possibly frustrating the interpretation of non-detects.

Another way that species and, sometimes strains, can be effectively detected and monitored in metagenomic surveys is by performing read recruitment plots (Fig. 2). In these plots, the reads of a metagenome are mapped against the genome representing the population with a read mapping tool (Boratyn et al., 2019), and the mapping patterns can reveal sequence discontinuities and gene content diversity (Konstantinidis and DeLong, 2008; Rodriguez-R and Konstantinidis, 2016; Rusch et al., 2007). Therefore, read recruitment plots can provide a transparent and quantitative view of the natural population in a sample, which can be highly useful for several downstream analyses. For instance, radical changes in the sequencing depth of specific regions of a genome by reads from timeseries metagenomes may signify genes gained or lost by the population due to selection for or against the corresponding functions, respectively, by the prevailing conditions during sampling (Bendall et al., 2016; Meziti et al., 2019). Yet, this approach trades throughput and speed for granularity and transparency of read mapping results of only a single genome (or gene) of interest at a time. Certainly, this framework is advantageous for metagenomic use cases like pathogen surveillance but cumbersome elsewhere. We see taxonomic profilers and read recruitment plotting as complementary approaches, i.e., deploying read recruitment plotting for a few species of interest while the community composition can be ascertained via the best performing taxonomic profiler for a given dataset.

Recently, our team has advanced the read recruitment plot tool to provide additional information based on read mapping results. This information includes what is the average sequencing depth of the reference genome, the average identity of the mapped reads to the reference, which we denote as ANIr (i.e., ANI based on reads), and whether related species exist in the sample. A couple use cases are shown in Fig. 2 and additional cases are documented in our recent publication (Gerhardt et al., 2021). ANIr is an important population genetics parameter to estimate because it reflects the clonality of the sampled population of the species when the reference genome is a good representative of the population. For example, if ANIr changes, this could indicate the emergence of adapted genotypes or strain replacement that could be targeted to identify the adaptive genes in future studies. If the reference genome is not a good representative of the population sampled by the metagenomic dataset (e.g., not many high identity reads evenly map to it across its total sequence), then ANIr reflects the level of divergence of the reference from the population (Fig. 2). In such cases, and also cases where multiple closely related populations exist in the sample, it is important to perform competitive read mapping against representatives of each (related) population(or species) for more accurate read mapping and robust estimations of relative abundance, as we described recently (Meziti et al., 2021; Viver et al., 2021). Therefore, read recruitment plots can be used to detect a target feature such as a gene or genome in a metagenome and characterize its allelic diversity. Further, the new read recruitment plots have been made interactive such that users are able to point the mouse cursor to a specific gene or region in the reference genome or a read and obtain information on the functional annotation of the corresponding sequences and other associated metadata, view multiple reference genomes/genes at a time, and more.

5. Metagenomics-based estimation of relative abundance and limit of detection

As noted above, methods for estimating relative abundance vary greatly. A common approach utilized across many manual and automated workflows is to estimate abundance as the approximate sequencing depth based on the number of reads (or base pairs) mapped to a reference genome (or gene) divided by the respective total number of reads (or base pairs) in the sample. While this is a straightforward way

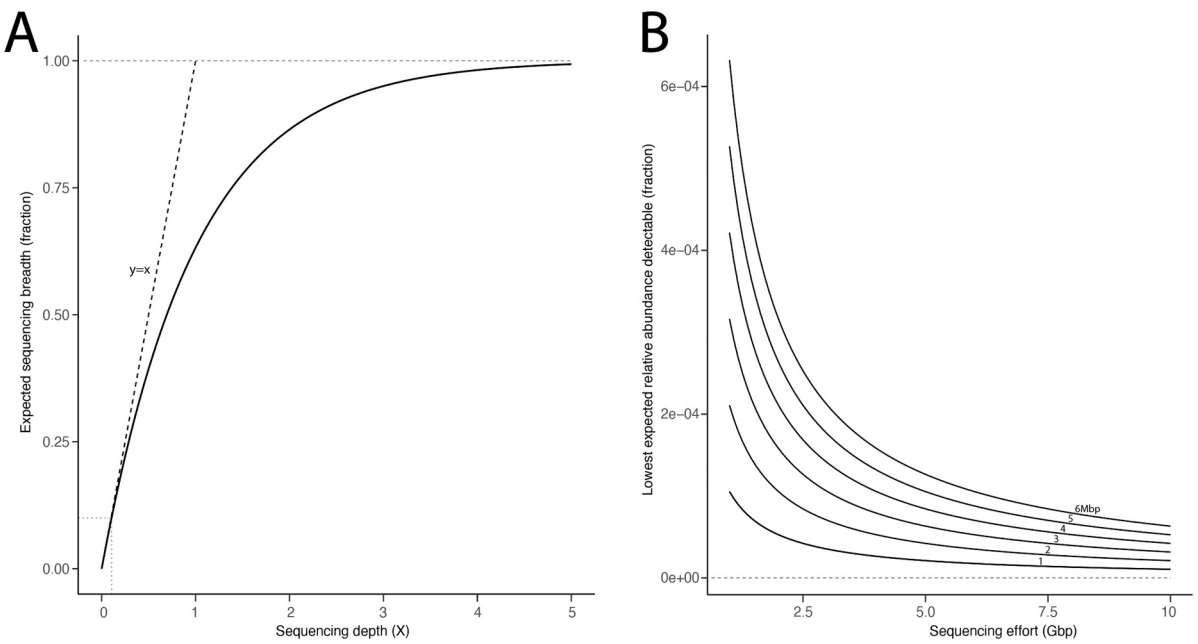


Fig. 3. Theoretical estimates of sensitivity for metagenomic datasets based on sequencing effort, depth, and breadth expectations. Panel A shows the relationship between expected sequencing breadth and sequencing depth proposed by Lander and Waterman (1988). Note that for low sequencing breadths the expectation of producing overlapping sequence reads is low as shown by the nearly linear relationship at this level. Both the proposed general purpose sequencing breadth detection limit of 10 % and the corresponding sequencing depth [−ln(0.9)] are marked for convenience. Panel B relates this information back to the relative abundance of a target and estimates of the lowest relative abundance that is expected to be detectable based on the amount of sequencing effort applied (in Gbp) and target's genome size (1-6Mbp genome sizes shown).

for communicating relative abundance, it produces values that likely correspond to the fraction of DNA belonging to the target and not the fraction of the community occupied by the target. Alternatively, a simple average of sequencing depth, without a clear denominator for normalizing the value to sequencing effort, is sometimes used. Lastly, relative abundance communicated as "reads per kilobase per million mapped reads" (RPKM) has also been commonly used. However, these measures can be prone to both false-positive read mappings due to unusual gene features as well as issues of relative abundance estimation caused by differences in the average genome size of the sampled microbial communities. To precisely estimate the limit of detection for a target genome in a metagenome, and thus help calculate its relative abundance, we recently developed and validated the imGLAD algorithm (*in-silico* metagenomes for Genomic Low-Abundance Detection) based on inoculation of known cell concentrations of *E. coli* on plant leaves followed by metagenome sequencing (Castro et al., 2018). Key observations from the imGLAD work are: i) that almost all genomes have regions (most often, genes) that are not reliable for diagnostics, such as highly conserved rRNA genes and recently horizontally transferred genes. These regions/genes may (misleadingly) recruit reads from co-occurring relatives in environmental samples, even when the actual target genome is absent, and ii) that when about 10 % of a genome recruits reads, it is generally enough sequencing breadth to sidestep the abovementioned limitation and provide reliable detection.

As a result of the above, we suggest replacing the simple average sequencing depth for reporting relative abundance with a more sophisticated metric that we term TAD-80 (truncated average sequencing depth over the middle 80 % of indices sorted by depth). TAD-80 is an average of sequencing depths calculated over only the middle 80 % of depths across the whole reference sequence, which is typically a whole genome. In detail, TAD-80 calculation involves sorting the base pair positions of the reference sequence by decreasing sequencing depth and then averaging across only the positions (or indices) remaining after removal of the top 10 % and bottom 10 % of positions. This approach automatically ensures nonzero values are reported only for genomes with sequencing breadth above 10 %, meeting (or exceeding) our suggested limit of detection from (ii) above (Rodriguez et al.,2020). Importantly, the removal of positions with high sequencing depth removes the problematic, non-diagnostic genes mentioned above that recruit spurious matches from co-occurring relatives in a sample (Castro et al., 2018). Further, the exclusion of high depth positions alone would likely result in a systematic underestimation of sequencing depth for several genomes, but the removal of an equal proportion of low depth positions counterbalances that to ensure that the TAD-80 value remains unbiased. To further facilitate the use of TAD-80, the new read recruitment plot tool automatically calculates TAD-80 values for a wide range of nucleotide identity cutoffs for mapping reads (default is 95 % identity) that may be used to describe the sequence-discrete boundaries of a species, based on the level of intra-population diversity presented on the same plot.

While the use of TAD-80 improves upon simple average sequencing depth for estimating relative abundance, we further suggest normalizing sequencing depth by prokaryotic genome equivalents (GEQ) in the sample (Nayfach and Pollard, 2015) to arrive at a final estimate of the relative abundance of a genome. The use of GEQ essentially controls for genome size differences among samples or microbial communities and is generally a more robust metric than simpler alternatives, most notably RPKM (Nayfach and Pollard, 2016). Though different metrics could provide very similar estimates if the average genome sizes of the communities compared are similar e.g., timeseries datasets of the same community (Konstantinidis et al., 2009; Nayfach and Pollard, 2015; Rodriguez et al., 2020). Specifically, a microbial community that is enriched in organisms with larger genome sizes will harbor the target genome in higher relative abundance compared to a community enriched in smaller genome sizes (smaller sequencing space available) when the target shows same sequencing breadth and depth in datasets from the two communities. Normalizing TAD-80 by GEQ results in a final estimate of relative abundance that intrinsically measures the abundance of a target genome in units of community fraction (i.e., percent of total genomes or cells as opposed to percentage of DNA) in a

manner more appropriate for direct abundance comparisons across communities or samples. Continual refinement of the reference universal gene sets used for GEQ estimation, separate for bacterial and eukaryotic fractions and facilitated by the continuous increase in the number of complete reference genomes, will further improve the accuracy of this approach in the future (Lind and Pollard, 2021).

To summarize: we propose the use of TAD-80 values higher than zero as a general purpose threshold for determining the limit of detection of a sequencing effort in detecting a target genome/sequence of interest in a metagenomic dataset, and to use such non-zero TAD-80 values normalized by GEQ as the final estimate of (non-zero) relative abundance in units of community fraction. Note, however, that even after these normalizations, the resulting abundance data represents only relative abundance (e.g., % of total community or cells). If the goal is to obtain absolute abundances (number of copies per sample volume) then additional work is necessary. For example, absolute abundances can be obtained by spiking in reference DNA (or cells for the step prior to DNA extraction) into the sample at known concentrations as an internal standard, as suggested previously (Poretsky et al., 2005). The limit of detection for a certain amount of metagenomic sequencing effort can then be estimated based on the relationships between absolute (or relative) abundance of the target, the chosen detection criteria, and sequencing effort, as described recently (Lindner et al., 2022) and summarized here for convenience (Table 1). We provide a visual representation of this sort of general purpose LOD for metagenomic work in Fig. 3 based on previous work (Castro et al., 2018; Lindner et al., 2022; Wendl et al., 2013). Although other solutions exist, this framework is based on one of the early relationships proposed between expected sequencing breadth and sequencing depth for a target genome in an experiment (Lander and Waterman, 1988; Wendl et al., 2013; Fig. 3). Though, it is especially critical to note that this sort of theoretical framework is not an appropriate substitution for spike-in/process controls, mock microbial standards, etc. which assist researchers in determining the actual sensitivity of their metagenomic experiments (Crossette et al., 2021; Sczyrba et al., 2017). Further, this theoretical LOD could be conservative in identifying a target organism as present in a metagenome, and thus users could visually inspect the corresponding recruitment plots to make a call about presence in cases where the relative abundance of the target is at or just below this LOD and higher sensitivity is required.

This theoretical framework can also be useful in establishing expectations for planning estimates of the sequencing effort necessary for studies aimed at surveilling pathogens via the equations described in Table 1. These equations theorize the sensitivity of a metagenome based on the minimum sequencing depth (ρ_{LOD}) chosen to call confident detection via the relationship theorized for relating sequencing breadth and sequencing depth at low values for both (Lander and Waterman, 1988; Fig. 3). These equations can then estimate either the minimum sequencing effort (S_{min}) needed to detect a population given some *a priori* knowledge about its in situ relative abundance or communicate the smallest detectable population size (C_{min}) based on the number of observed GEQ and average genome size for the community. For the limit of detection, the minimum sequencing breadth (and consequently ρ_{LOD}) could be adjusted if viewed necessary e.g., genomes undergoing infrequent horizontal gene transfer may be reliably detectable even with lower breadth thresholds than 10 % (Castro et al., 2018).

In our experience with human fecal and freshwater samples that are medium-to-high complexity (Rodriguez-R and Konstantinidis, 2014), the limit of detection of a 5 gigabases (or Gbp) sequencing effort for *Escherichia coli* or similar organisms is at about 0.01 %–0.001 % of the total microbial community (percent of total cells or genomes). That is, such sequencing efforts have a range of detection (and abundance estimation) of about 5 orders of magnitude (from 100 % to 0.001 % of the total). There are additional factors capable of confounding estimates of relative abundance. These factors include various sources of error such as substantial proportions of eukaryotic and viral sequences, G + C% bias caused by amplification-based library preparation methods, bias introduced against certain physiologies due to differing cell lysis efficiencies, etc. Viral sequences do not generally contribute to GEQ estimation because these genomes do not usually carry universal genes (except the giant viruses (Schulz et al., 2020)), and thus do not affect relative abundance normalization by GEQ but affect – for example – RPKM estimations. In cases that significant differences are observed, it is advisable to remove the eukaryotic sequences (e.g., by read mapping against NCBI genomes) prior to the genome equivalent estimation step or estimation of relative abundance. Currently, there exists no universal methodology for controlling relative abundance biases in bioinformatic analysis of metagenomic data and researchers should evaluate their individual use cases to decide the best approach for their work (Lin and Peddada, 2020).

6. Identification of the etiological agent of a disease based on metagenomic datasets

Identification of the etiological agent (pathogen) of a disease could be a challenging task, especially in cases where defining features of a pathogen are unclear (reviewed in Denamur et al., 2021). Such cases include instances where virulence factors are not known (e.g., *Salmonella* spp.) and/or some of the known virulence factors (e.g., iron acquisition genes or adhesins) are frequently shared with commensal close relatives. For instance, enteric infections caused by diarrheagenic strains of *Escherichia coli* (DEC) represent such challenging cases to diagnose, despite their apparent (high) frequency and importance for children mortality, especially in low-income countries (Collaborators, 2018). We recently developed an integrated approach, using the principles mentioned above, to identify the etiological agent of diarrheal disease based on metagenomic datasets obtained from fecal samples, thus facilitating the diagnosis of cases of DEC and other etiological agents (Pena-Gonzalez et al., 2019). The approach combines three key data features to determine the etiological agent: i) the in-situ metagenomic abundance of the suspected pathogen(s) should be higher in the disease vs. control samples, ii) the level of clonality of the pathogen population should be higher (less intra-population diversity or higher ANIr) when causing disease compared to commensal relatives in the same disease sample or in control samples and, iii) the key virulence genes, when known, should be detected/present in the metagenome, even if they are not necessarily assembled as part of the MAG that represents the pathogen, and should be at comparable sequencing depths to that of the pathogen genome or MAG, unless carried on multicopy plasmids. The virulence genes are frequently carried on plasmids or other mobile elements that are not assembled as part of a MAG, especially when short-read data are used (Meziti et al., 2021).

Using this strategy, we have recently elucidated the causative agents of foodborne outbreaks caused by DEC (Pena-Gonzalez et al., 2019) and *Salmonella enterica* (Huang et al., 2017) strains and showed that our integrated approach provides resolution and diagnostic signatures that are not attainable by the traditional, culture-based approaches. Most notably, our approach can elucidate the distinct signatures of different enteric pathogens on the gut microbiome, which could be useful for diagnostics on its own, and also provides the level of intra-population sequence and gene-content diversity for the detected pathogen (Huang et al., 2017; Pena-Gonzalez et al., 2019). Notably, in at least 1/3 of the diarrheal cases examined that involved DEC infections in Northern Ecuador, this metagenomic approach identified a different etiological agent compared to the traditional approach based on isolation followed by PCR typing for the pathogen-diagnostic genes (Pena-Gonzalez et al., 2019). In most of these cases, it appeared that the isolation-based approach recovered a pathogen that was a minor player of the microbial community (rare biosphere), as opposed to the dominant pathogen, a known limitation of culture-based approaches that can recover a (target) organism present in a sample even as a single cell (but unlikely to cause disease at such low abundance). Further, the metagenomic

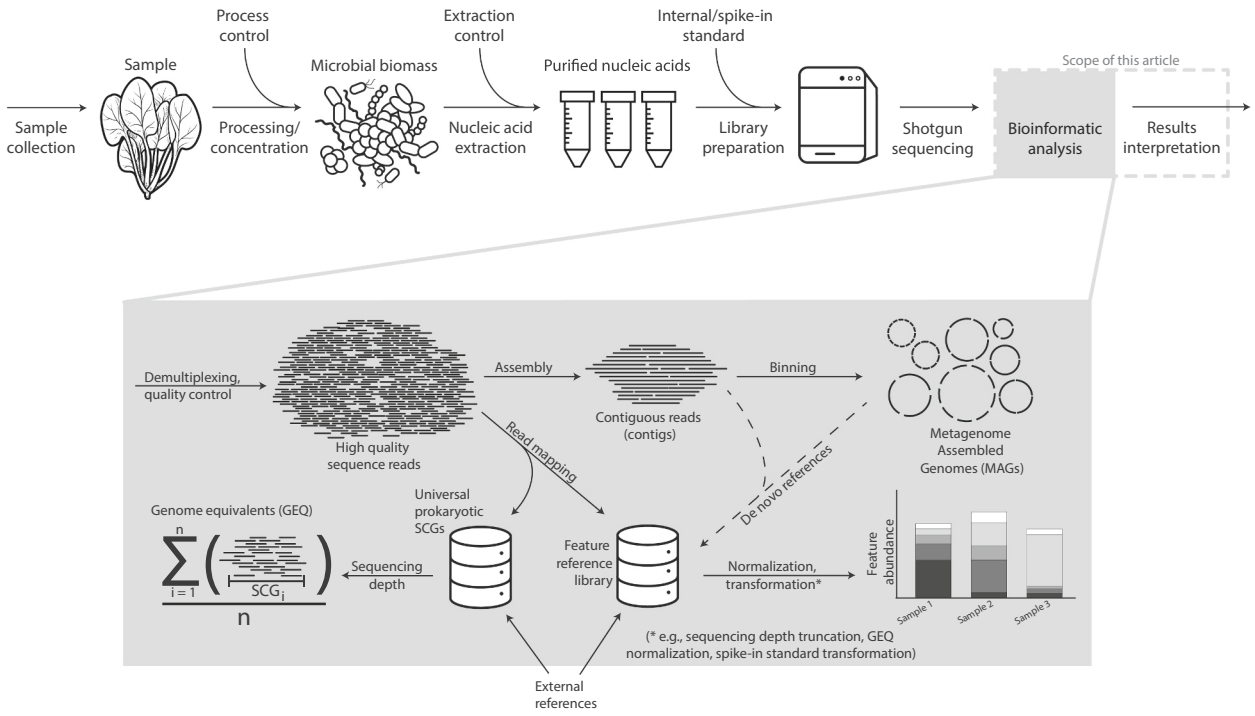


Fig. 4. A schematic representation of the shotgun metagenomic approach for pathogen detection in a sample. Top: the typical wet-lab steps for obtaining a metagenome, highlighting where controls and internal standards may be introduced. Bottom: the dry-lab steps for determining the abundance of a target feature (genome or gene) in the metagenome. Targets may be among the feature reference library or make up the library exclusively. Abundances produced for feature reference sequences are relative abundances unless appropriate controls or internal (spike-in) standards have been included in the previous wet-lab steps. The common normalization and transformation approaches for abundance estimation are noted within the figure and reviewed at length within the text. The cumulative abundances of features (bottom right) are purposefully shown not to sum to one to illustrate incomplete capture of the metagenomic data even with combined external and de novo (internal) reference sequences, which is often observed in practice. Single copy gene is abbreviated as SCG for genome equivalent calculation.

approach can be significantly faster than the traditional approach, e.g., <1 day vs. at least 2–3 days, which is important for diagnostic and treatment actions. Therefore, we suggest complementing traditional, culture-based approaches with a metagenomic approach like the one described above for clinical samples as well as food samples that are challenging to process with the traditional approach and/or when higher resolution or forensic/tracking evaluation is needed.

7. Challenges remaining for metagenomics approaches in detecting target organisms

As reviewed herein, metagenomics involves neither the isolation nor the enrichment of target organisms. Thus, despite the advancements mentioned above and its relatively high throughput, metagenomics is inherently an untargeted approach, and sequencing microbial communities where the interesting feature(s) are rare can be quite expensive. That is, the associated cost can still be prohibitive for processing a large number of samples on a routine basis especially if the target(s) are anticipated to be at relatively low abundances (e.g., <0.001 % of the total community). At the time of this writing, an "average" metagenomic sample (e.g., 5 Gbp/sample) costs about a couple hundred US dollars, including DNA extraction, library creations and sequencing costs. Perhaps more importantly, often highly trained personnel and substantial computational resources are required for the analysis of the resulting sequence data and interpretation of results. Moreover, the limit of detection is still a couple orders of magnitude – or more – higher than that of PCR-based methods, although the exact difference depends on the relative abundance of the target organism and the level of the sequencing effort applied. In our experience and using the equations shown in Table 1, metagenomic detection of a spiked-in pathogen onto plant leaves, which are characterized by relatively simple and low-biomass microbial communities, was possible even when the target was inoculated at ~80 cells per gram of sample (Castro et al., 2018). For a typical fecal sample and a sequencing effort around 3–5 Gbp/sample, this limit of detection is likely a couple orders of magnitude higher, at 10,000 to 100,000 cells or about 0.01 % to 0.001 % of the total microbial community assuming that usually 10^9 to 10^{10} cells are sampled with the typical sample volumes used (0.2–0.5 g) and 10 % sequencing breadth is required for robust detection. The latter limit of detection is significantly higher than that which is usually associated with PCR-based approaches. It should be noted, however, that the biomass of fecal samples and their DNA yields can be highly variable, and thus their actual limit of detection could be an order of magnitude or more different than the values mentioned above. The approach outlined in Table 1 and Fig. 3 can be used to estimate limits of detection for various situations based on the relative abundance of the target organism, sequencing effort applied, and sequence breadth threshold used for detection.

Another notable limitation of metagenomics that is nonetheless shared with other culture-independent techniques including PCR, is that it cannot easily distinguish between living and dead cells or naked DNA. This issue has been discussed extensively in the literature and there are approaches to selectively remove dead cells (e.g., application of propidium monoazide or PMA) or naked DNA (e.g., by a DNAse treatment) but typically require additional steps and protocol optimization for the sample matrix of interest (Nocker et al., 2006). Accordingly, these approaches have not been met with wide acceptance yet. Further, in our experience and unless there has been a recent application of a chemical (e.g., antibiotic) or processing (e.g., heat exposure) that killed the microbial cells in a sample, species that are detected at high relative abundances by metagenomics (e.g., >0.1–0.01 % of the total community) represent alive and active members of the community. Naked DNA or dead cells are typically represented by only a few reads per species (e.g., members of the rare biosphere) although, altogether, such sequences could make up a large fraction of the total community, even close to 50

% of the total for some soil and bioaerosols (dust) samples (Sogin et al., 2006). In cases that this is important, culture-independent approaches should be combined with cultivation to ensure that the detected targets are alive, and thus verify that they pose a public (or other) health risk. In fact, having the genome of the isolate(s) derived from the sample available can greatly facilitate the approaches mentioned above, e.g., it can serve as the reference sequence to perform the read-recruitment plot against and quickly assess if the genome is detectable in the sample, determine its relative abundance as a fraction of the total community, and elucidate how well it represents -or not- the natural population in the sample. While the focus on this review was bacterial pathogens, we expect that the approaches described herein and summarized in Fig. 4 can be used for viral and eukaryotic pathogens, albeit with some optimizations for the complexity of the genomes targeted. For instance, eukaryotic genomes are generally engaged in less frequency horizontal transfer than bacterial genomes, and thus the 10 % sequencing breadth for reliable detection could be too high for such genomes.

Declaration of competing interest

The authors declare no competing interest.

Data availability

all data used are already publicly available at NCBI and EBI databases.

Acknowledgments

This work has been supported by the US National Science Foundation (Award No 1759831 and 2129823) and the US EPA (Award No 84020301-0).

References

Beghini, F., McIver, L.J., Blanco-Miguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A.M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E.A., Segata, N., 2021. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 10.

Bendall, M.L., Stevens, S.L., Chan, L.K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S.G., Bertilsson, S., Moran, M.A., Shade, A., Newton, R.J., McMahon, K.D., Malmstrom, R. R., 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 10, 1589–1601.

Benson, D.A., Karsch-Mizrachi, I., Clark, K., Lipman, D.J., Ostell, J., Sayers, E.W., 2012. GenBank. Nucleic Acids Res. 40, D48–D53.

Blanco-Miguez, A., Beghini, F., Cumbo, F., McIver, L.J., Thompson, K.N., Zolfo, M., Manghi, P., Dubois, L., Huang, K.D., Thomas, A.M., Nickols, W.A., Piccinno, G., Piperni, E., Puncochar, M., Valles-Colomer, M., Tett, A., Giordano, F., Davies, R., Wolf, J., Berry, S.E., Spector, T.D., Franzosa, E.A., Pasolli, E., Asnicar, F., Huttenhower, C., Segata, N., 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nat. Biotechnol. 41, 1633–1644.

Boratyn, G.M., Thierry-Mieg, J., Thierry-Mieg, D., Busby, B., Madden, T.L., 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinform. 20, 405.

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

Castro, J.C., Rodriguez, R.L., Harvey, W.T., Weigand, M.R., Hatt, J.K., Carter, M.Q., Konstantinidis, K.T., 2018. imGLAD: accurate detection and quantification of target organisms in metagenomes. PeerJ 6, e5882.

Chen, L.X., Anantharaman, K., Shaiber, A., Eren, A.M., Banfield, J.F., 2020. Accurate and complete genomes from metagenomes. Genome Res. 30, 315–333.

Collaborators, G. B. D. L. R. I, 2018. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Infect. Dis. 18, 1191–1210.

Crossette, E., Gumm, J., Langenfeld, K., Raskin, L., Duhaime, M., Wigginton, K., 2021. Metagenomic quantification of genes with internal standards. mBio 12.

Denamur, E., Clermont, O., Bonacorsi, S., Gordon, D., 2021. The population genetics of pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 19, 37–54.

Dijkshoorn, L., Ursing, B.M., Ursing, J.B., 2000. Strain, clone and species: comments on three basic concepts of bacteriology. J. Med. Microbiol. 49, 397–401.

Doolittle, W.F., 2019. Speciation without species: a final word. Philos. Theory Pract. Biol. 11.

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., Hanage, W.P., 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science 323, 741–746.

Gerhardt, K., Ruiz-Perez, C.A., Rodriguez, R.L., Conrad, R.E., Konstantinidis, K.T., 2021. RecruitPlotEasy: an advanced read recruitment plot tool for assessing metagenomic population abundance and genetic diversity. Front. Bioinform. 1, 826701.

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biol. 3, 6.

Handelsman, J., Tiedje, J., Alvarez-Cohen, L., Ashburner, M., Cann, I., Delong, E., Doolittle, W., Fraser-Liggett, C., Godzik, A., Gordon, J., Riley, M., Schmidt, T., 2007. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. The National Academies Press, Washington, DC.

Huang, A.D., Luo, C., Pena-Gonzalez, A., Weigand, M.R., Tarr, C.L., Konstantinidis, K.T., 2017. Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods. Appl. Environ. Microbiol. 83.

Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 9, 5114.

Jones, M.B., Highlander, S.K., Anderson, E.L., Li, W., Dayrit, M., Klitgord, N., Fabani, M. M., Seguritan, V., Green, J., Pride, D.T., Yooseph, S., Biggs, W., Nelson, K.E., Venter, J.C., 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc. Natl. Acad. Sci. U. S. A. 112, 14024–14029.

Knoppel, A., Knopp, M., Albrecht, L.M., Lundin, E., Lustig, U., Nasvall, J., Andersson, D. I., 2018. Genetic adaptation to growth under laboratory conditions in *Escherichia coli* and *Salmonella enterica*. Front. Microbiol. 9, 756.

Konstantinidis, K.T., DeLong, E.F., 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2, 1052–1065.

Konstantinidis, K.T., Braff, J., Karl, D.M., DeLong, E.F., 2009. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. Appl. Environ. Microbiol. 75, 5345–5355.

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

Lawrence, J.G., 2002. Gene transfer in bacteria: speciation without species? Theor. Popul. Biol. 61, 449–460.

Lin, H., Peddada, S.D., 2020. Analysis of compositions of microbiomes with bias correction. Nat. Commun. 11, 3514.

Lind, A.L., Pollard, K.S., 2021. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. Microbiome 9, 58.

Lindner, B.G., Suttner, B., Zhu, K.J., Conrad, R.E., Rodriguez, R.L., Hatt, J.K., Brown, J., Konstantinidis, K.T., 2022. Toward shotgun metagenomic approaches for microbial source tracking sewage spills based on laboratory mesocosms. Water Res. 210, 117993.

Lu, J., Salzberg, S.L., 2020. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. Microbiome 8, 124.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., Konstantinidis, K.T., 2011. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc. Natl. Acad. Sci. U. S. A. 108, 7200–7205.

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R.J., Gevers, D., 2015. ConStrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol. 33, 1045–1052.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. 95, 3140–3145.

McLaren, M.R., Willis, A.D., Callahan, B.J., 2019. Consistent and correctable bias in metagenomic sequencing experiments. Elife 8.

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7, 11257.

Meziti, A., Tsementzi, D., Rodriguez, R.L., Hatt, J.K., Karayanni, H., Kormas, K.A., Konstantinidis, K.T., 2019. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. ISME J. 13, 767–779.

Meziti, A., Rodriguez, R.L., Hatt, J.K., Pena-Gonzalez, A., Levy, K., Konstantinidis, K.T., 2021. 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. 87.

Morton, J.T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L.S., Edlund, A., Zengler, K., Knight, R., 2019. Establishing microbial composition measurement standards with reference frames. Nat. Commun. 10, 2719.

Nayfach, S., Pollard, K.S., 2015. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 16, 51.

Nayfach, S., Pollard, K.S., 2016. Toward accurate and quantitative comparative metagenomics. Cell 166, 1103–1116.

Nearing, J.T., Comeau, A.M., Langille, M.G.I., 2021. Identifying biases and their potential solutions in human microbiome studies. Microbiome 9, 113.

Nocker, A., Cheung, C.Y., Camper, A.K., 2006. Comparison of propidium monoazide with ethidium monoazide for differentiation of live vs. dead bacteria by selective removal of DNA from dead cells. J. Microbiol. Methods 67, 310–320.

Olm, M.R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P.B., Banfield, J.F., 2020. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. mSystems 5.

Olm, M.R., Crits-Christoph, A., Bouma-Gregson, K., Firek, B.A., Morowitz, M.J., Banfield, J.F., 2021. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat. Biotechnol. 39, 727–736.

Parker, C.T., Tindall, B.J., Garrity, G.M., 2019. International code of nomenclature of prokaryotes. Int. J. Syst. Evol. Microbiol. 69 (1A), S1–S111.

Pena-Gonzalez, A., Soto-Giron, M.J., Smith, S., Sistrunk, J., Montero, L., Paez, M., Ortega, E., Hatt, J.K., Cevallos, W., Trueba, G., Levy, K., Konstantinidis, K.T., 2019. Metagenomic signatures of gut infections caused by different *Escherichia coli* pathotypes. Appl. Environ. Microbiol. 85.

Poretsky, R.S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M., Pate, W. M., Moran, M.A., Hollibaugh, J.T., 2005. Analysis of microbial gene transcripts in environmental samples. Appl. Environ. Microbiol. 71, 4121–4126.

Rodriguez, et al., 2020. Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. Environ. Microbiol. https://doi.org/10.1111/1462-2920.15112.

Rodriguez, R.L., Jain, C., Conrad, R.E., Aluru, S., Konstantinidis, K.T., 2021. Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species". Nat. Commun. 12, 4060.

Rodriguez-R, L.M., Konstantinidis, K.T., 2014. Estimating coverage in metagenomic data sets and why it matters. ISME J. 8, 2349–2351.

Rodriguez-R, L.-M., Konstantinidis, K.T., 2016. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints 4, e1900v1. https://doi.org/10.7287/peerj.preprints.1900v1.

L. M. Rodriguez-R, R. E. Conrad, T. Viver, D. J. Feistel, B. G. Lindner, S. N. Venter, L. H. Orellana, R. Amann, R. Rossello-Mora, Konstantinos T. Konstantinidis. An ANI gap within bacterial species that advances the definitionsof intra-species units. mBio, In press. Doi: 10.1128/mbio.02696-23.

Rossello-Mora, R., Amann, R., 2015. Past and future species definitions for bacteria and archaea. Syst. Appl. Microbiol. 38, 209–216.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcon, L.I., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C., 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5, e77.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., Woyke, T., 2020. Giant virus diversity and host interactions through global metagenomics. Nature 578, 432–436.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jorgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvociute, M., Hansen, L.H., Sorensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.H., Liao, Y.C., Silva, G.G.Z., Cuevas, D. A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.P., Goker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A. E., Rattei, T., McHardy, A.C., 2017. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. Nat. Methods 14, 1063–1071.

Seabolt, M.H., Konstantinidis, K.T., Roellig, D.M., 2021. Hidden diversity within common protozoan parasites as revealed by a novel genomotyping scheme. Appl. Environ. Microbiol. 87.

Sheppard, S.K., McCarthy, N.D., Falush, D., Maiden, M.C., 2008. Convergence of Campylobacter species: implications for bacterial evolution. Science 320, 237–239.

Simmonds, P., Adams, M.J., Benko, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., Hull, R., King, A.M., Koonin, E.V., Krupovic, M., Kuhn, J.H., Lefkowitz, E.J., Nibert, M.L., Orton, R., Roossinck, M.J., Sabanadzovic, S., Sullivan, M.B., Suttle, C.A., Tesh, R.B., van der Vlugt, R.A., Varsani, A., Zerbini, F.M., 2017. Consensus statement: virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 15, 161–168.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., Herndl, G.J., 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc. Natl. Acad. Sci. U. S. A. 103, 12115–12120.

Ursing, J.B., Rossello-Mora, R.A., Garcia-Valdes, E., Lalucat, J., 1995. Taxonomic note: a pragmatic approach to the nomenclature of phenotypically similar genomic groups. Int. J. Syst. Evol. Microbiol. 45, 604.

van Dijk, L.R., Walker, B.J., Straub, T.J., Worby, C.J., Grote, A., Schreiber, H.L.t., Anyansi, C., Pickering, A.J., Hultgren, S.J., Manson, A.L., Abeel, T., Earl, A.M., 2022. StrainGE: a toolkit to track and characterize low-abundance strains in complex microbial communities. Genome Biol. 23, 74.

Van Rossum, T., Ferretti, P., Maistrenko, O.M., Bork, P., 2020. Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. 18, 491–506.

Viver, T., Conrad, R.E., Orellana, L.H., Urdiain, M., Gonzalez-Pastor, J.E., Hatt, J.K., Amann, R., Anton, J., Konstantinidis, K.T., Rossello-Mora, R., 2021. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. ISME J. 15 (4), 1178–1191. https://doi.org/10.1038/s41396-020-00842-5.

Viver, T., Conrad, R.E., Rodriguez-R, L.M., Ramirez, A.S., Venter, S.N., Rocha-Cardenas, J., Segura, M.L., Amann, A., Konstantinidis, K.T., Rossello-Mora, R., 2023. Towards estimating the number of strains that make up a natural bacterial population. bioRxiv.

Wendl, M.C., Kota, K., Weinstock, G.M., Mitreva, M., 2013. Coverage theories for metagenomic DNA sequencing based on a generalization of Stevens' theorem. J. Math. Biol. 67 (5), 1141–1161. https://doi.org/10.1007/s00285-012-0586-x.

Yan, Y., Nguyen, L.H., Franzosa, E.A., Huttenhower, C., 2020. Strain-level epidemiology of microbial communities and the human microbiome. Genome Med. 12, 71.