Annual Review of Microbiology

What Is Metagenomics Teaching Us, and What Is Missed?

Felicia N. New and Ilana L. Brito

Meinig School of Biomedical Engineering, Cornell University, Ithaca, New York 14853, USA

Annu. Rev. Microbiol. 2020. 74:117–35

The *Annual Review of Microbiology* is online at micro.annualreviews.org

https://doi.org/10.1146/annurev-micro-012520-072314

Keywords

microbiome, metagenomics, DNA sequencing, microbiology, computational biology

Abstract

Shotgun metagenomic sequencing has revolutionized our ability to detect and characterize the diversity and function of complex microbial communities. In this review, we highlight the benefits of using metagenomics as well as the breadth of conclusions that can be made using currently available analytical tools, such as greater resolution of species and strains across phyla and functional content, while highlighting challenges of metagenomic data analysis. Major challenges remain in annotating function, given the dearth of functional databases for environmental bacteria compared to model organisms, and the technical difficulties of metagenome assembly and phasing in heterogeneous environmental samples. In the future, improvements and innovation in technology and methodology will lead to lowered costs. Data integration using multiple technological platforms will lead to a better understanding of how to harness metagenomes. Subsequently, we will be able not only to characterize complex microbiomes but also able to manipulate communities to achieve prosperous outcomes for health, agriculture, and environmental sustainability.

Contents

INTRODUCTION 118
WHAT HAS METAGENOMICS TAUGHT US? 119
A Trove of New (Draft) Genomes 119
Methods to Assemble Metagenomes 120
Metagenomic Assembly Quality 121
Better Compositional Data Than Taxonomic Profiling Alone 122
The Diversity of Host-Associated Viruses, Fungi, and Archaea 122
Host-Associated Phenotypes 123
Strain-Level Analyses 124
Functional Assessments of Metagenomes 124
Interesting Miscellanea 125
Horizontal Gene Transfer 125
Replication Rates of Organisms 126
WHAT HAS METAGENOMICS MISSED? 126
Ecoevolutionary Modeling 127
Phenotype and Comprehensive Functional Profiling 127
Spatial Analyses of the Microbiome 128
INNOVATION AND FUTURE DIRECTIONS 129

INTRODUCTION

The tools of microbiology—microscopy, culturing, and genetic engineering—have allowed researchers to observe, grow, and experiment on a small number of well-studied organisms, revealing insights into their biological, ecological, and evolutionary capacities. Yet microbes live nearly everywhere on Earth, have vast influence over ecosystem services and host health, and are dominant members of all three domains of life. Despite scientific awareness of this diversity, it has remained unexplored until recently. The advent of high-throughput sequencing platforms has rapidly enhanced our ability to understand the diversity of species in microbiomes by coupling physiological data with the underlying genetic data. Though metagenomes have given us glimpses into the diversity and function of complex microbiomes, this data can itself be incomplete, biased, and challenging. It is thus important to be extremely critical of and to understand the limitations of metagenomic data as the scientific community continues to embrace this technology.

Due to the cost, computational footprint, and analytical hurdles of whole-microbiome shotgun sequencing, amplicon sequencing of the 16S rRNA gene is used broadly to determine the taxonomic identities of members of a microbial community. The 16S rRNA gene, ubiquitous in all bacteria and archaea, was chosen as a genetic marker for taxonomic identification for several reasons. Barring some exceptions, this gene is evolutionarily stable, meaning that it has gone through little horizontal gene transfer, follows a molecular clock, and has regions of conservation and regions of divergence. For most microbiomes, several tens of thousands of sequences are adequate to assess the diversity in a sample (55), and as of 2020, the cost of DNA extraction, library preparation, and sequencing would cost less than $25–50 per sample, depending on the number and sample type, making this data type the most broadly accessible. With this accessibility has come streamlined analytical platforms, such as QIIME (12), UCHIME2 (35), mothur (95), and dada2 (24), using reference-based assignment of taxonomies and/or de novo sequence clustering.

Yet, sequencing this single gene to determine community composition results in several complications, arising from the facts that organisms may carry 1–15 genetically dissimilar, and occasionally fairly distant, copies; that artifacts, such as chimeras and jackpot effects, arise during the amplification process; and that the resolution of taxa varies depending on the branch of the bacterial tree of life. Despite innovations to solve these problems, both experimentally and computationally, with programs to remove chimeras and deal with errors inherent to the platform (24), significant biases may remain. As a result, the studies exploit the low cost and streamlined computational analyses of 16S rRNA data sets to cover either large dense time courses where increased coverage can mitigate the effects of noise or diverse ecologies where the differences are more robust.

With the increasing appreciation for dramatic phenotypic differences arising from differences in the genomic content of organisms and strains, there has been a movement toward using higher-resolution differences in 16S sequences, called amplicon sequence variants (ASVs). Previously, DNA sequences would be clustered at 97% sequence identity, a cutoff meant to distinguish between species while masking the effect of PCR (polymerase chain reaction) or sequencing errors. Among the benefits of using ASVs rather than clustered 16S sequences are a greater comparability across studies, greater reproducibility, and lack of reliance on previously curated reference libraries (23).

As part of this trend to obtain greater resolution, whole-microbiome shotgun sequencing, hereafter referred to as metagenomics, is becoming an important data type for many studies aiming to understand the mechanisms driving microbiome-associated traits. Rather than amplifying a single gene, all of the DNA within a sample is sequenced, regardless of whether it originated from bacteria. DNA is simply extracted, made into libraries, and sequenced either on a short-read platform (e.g., sequencing by synthesis, such as Illumina's platform) or on a long-read platform [e.g., single-molecule real-time (SMRT) sequencing, used by PacBio, or nanopore, used by Oxford Nanopore]. Recently, with the decreased costs of DNA sequencing and library preparation, studies have grown in breadth and scope.

This review focuses mainly on the benefits of using metagenomics and outlining the breadth of conclusions that can be made using currently available analytical tools, such as greater resolution of species and strains across phyla and functional content, while highlighting challenges of metagenomic data analysis (**Figure 1**). These major challenges include annotating function, given the relative lack of functional databases for environmental bacteria compared to model organisms, and the technical challenges of assembly and phasing in heterogeneous environmental samples.

WHAT HAS METAGENOMICS TAUGHT US?

A Trove of New (Draft) Genomes

Many of the earliest metagenomic studies used alignments to reference genomes to assess composition and function (62). Given a suitable reference catalog of genes, coding regions, or reference genomes, metagenomic data can be mapped to the reference using a typical alignment software. Yet, many environmental metagenomic samples, still to this day, lack appropriate representative reference genomes. A simple study where spores were selected from human gut microbiome samples revealed 45 novel candidate species (21), despite relatively deep study of the human gut microbiome. Advances in culturing of organisms are improving our reference libraries of previously unknown or unculturable species (65, 93), yet de novo assembly methods are still necessary to fill this knowledge gap. While single-genome assemblers were not appropriate for assembling metagenomes, because of the varying abundances of bacteria within a community, tools were

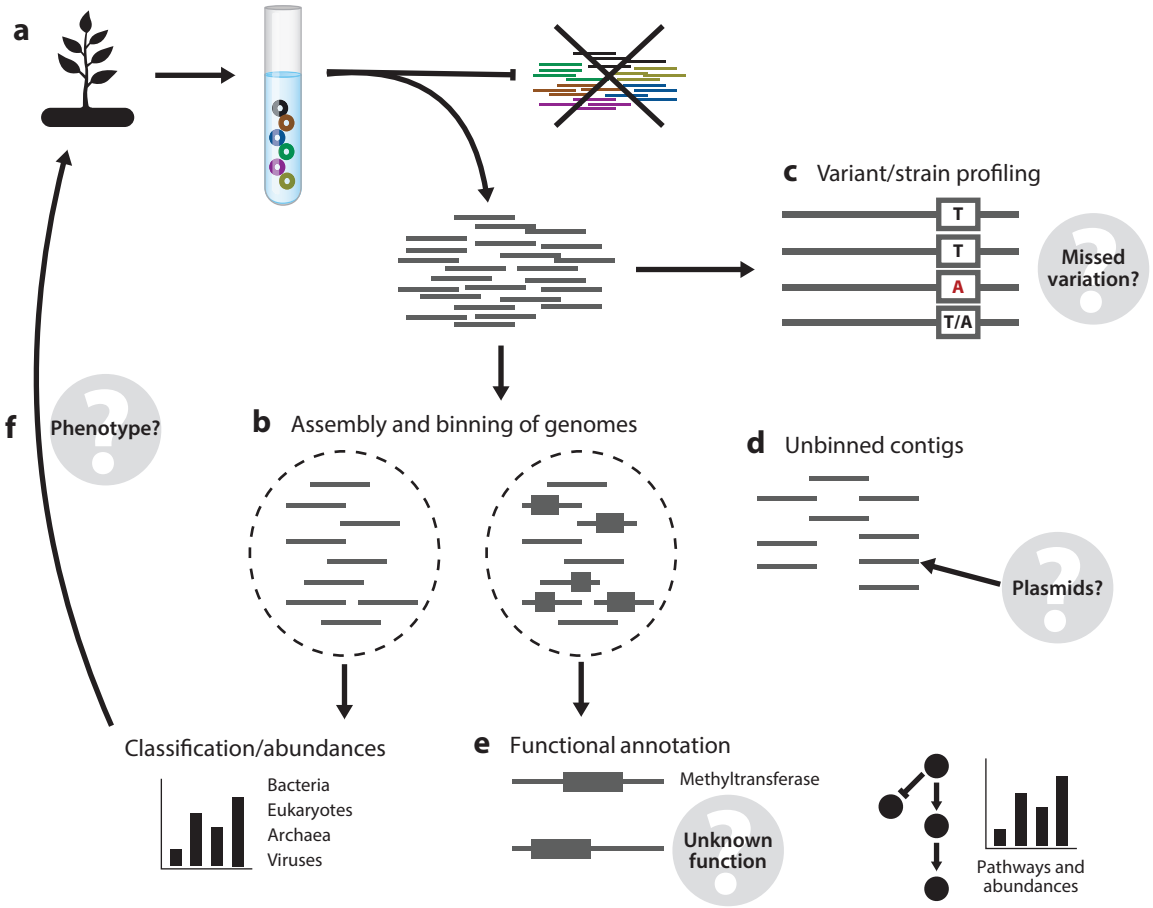

Figure 1

Metagenomic shotgun sequencing, the profiling of all DNA present within a microbiome sample, has many benefits and challenges. (*a*) DNA is extracted, made into libraries, and sequenced. The genomic and cellular context is lost during the process, and the output is reads that need to be organized into meaningful groups. (*b*) Assembly and binning are general steps for many analyses including taxonomic and functional profiling. However, many challenges and unknowns remain, here denoted by question marks: (*c*) Strain profiling methods have improved, but it is difficult to know what variation is missed; (*d*) unbinned contigs originating from plasmids, sequencing errors, or low-abundance (or low-sampled) organisms are indistinguishable; (*e*) many genes found in microbiome samples remain unannotated; and (*f*) it is difficult to link metagenomic data to host traits.

designed to account for and leverage these distinctive metagenomic qualities to assemble draft genomes within complex, heterogeneous assemblies.

Methods to Assemble Metagenomes

Overlap or consensus assembly methods were initially applied for DNA sequence assembly. These methods use a greedy algorithm and were originally created for Sanger sequencing (4). Given that pairwise comparisons of every read must be made, they are computationally expensive and became less suitable for next-generation sequencing (NGS). The algorithm underlying de Bruijn graph assemblers breaks the sequencing reads down into uniform k-mers of a specified size, k. The k-mers

are used as nodes in the graph, and overlapping nodes are connected by an edge. The assembler then constructs sequences based on the compiled graph. These methods (4, 101, 133) reduced the computational memory requirements because they essentially compresses the repetition inherent to NGS data, negating the need to perform pairwise read alignments. Despite the gains in performance, several challenges remain. As reads are broken down into k-mers, some genomic context is lost. Additionally, the choice of k-mer size and the choice of tools can significantly alter an assembly. One solution helpful for metagenomes has been to employ iterative de Bruijn graph assembly, which combines graphs from various k-mer sizes (78, 86). Currently, there is no single best practice. In the Critical Assessment of Metagenome Interpretation (CAMI) challenge (96) to assemble metagenomes, the simulated data assembly had 39,140 contigs and was 1.97 Gbp long. However, other programs resulted in a range of results: MEGAHIT (66) resulted in the largest assembly, with 587,607 contigs, of 1.91 Gbp, whereas the smallest assembly, produced by Ray Meta (11), was 12.3 Mbp long and had just 13,847 contigs. We recommend the review paper by Ayling et al. (4) and the CAMI challenge (96) for comparisons of these methods.

A subsequent challenge is making sense of the thousands of relatively small contigs that can result from assembling heterogeneous microbiome data. Whereas contigs could be binned into genomes based on taxonomic markers, genomes are typically fragmented, incomplete, and contaminated. New algorithms to bin contigs have led to higher-fidelity genome assembly. Binning algorithms use several different metrics to group contigs: DNA composition, GC content, tetranucleotide frequency, depth of sequencing coverage, and abundance or coabundance patterns across multiple samples (2, 53, 57, 127). An alternative method is to bin reads that are predicted to be derived from the same organism prior to assembly (28). There are also downstream tools that combine output from several binning tools (99, 107, 116) to refine and recombine contigs with the goal of identifying cleaner and more complete metagenome-assembled genomes (MAGs).

These techniques have been applied to large numbers of metagenomes to reconstruct draft genomes in a uniform manner. Nayfach et al. (77) reconstructed draft genomes from 3,810 publicly available human gut metagenomic samples, and Pasolli et al. (85) analyzed 9,428 metagenomes across microbiomes on different body sites. Both studies prioritized assemblies from international cohorts, including those in areas of the world in which fewer microbiome studies have been performed overall. The former study resulted in over 2,000 newly identified species, accounting for a 50% increase in the phylogenetic diversity of the human gut microbiome. Despite the creative use of recent tools to recover this vast diversity from individual gut samples, strain-level diversity and sequencing depth still pose a challenge to MAG assembly. The latter study reconstructed 154,723 MAGs, increasing the mappability of human metagenomic reads from around 67% to over 87% in gut microbiome samples used in this study, and from 65% to 82% in oral microbiome samples. These two studies highlight a particularly timely challenge to the field, the need to internationalize metagenomic sequencing efforts (88). Most of the unknown or uncharacterized species found in these studies were found in non-Western, low- and middle-income populations.

Metagenomic Assembly Quality

Following assembly, it is important to assess quality. Since there is often no ground truth in metagenomic sequencing for environmental samples, assembly quality is usually evaluated using summary statistics from single-genome assembly methods like size, contig N50, and maximum contig length (14). Completeness and contamination are the two main metrics that researchers rely on for assessing MAG quality. Completeness relies on the identification of marker gene sets and can miss strain heterogeneity. Likewise, contamination is derived from a set of single-copy marker genes and can be complicated if genes overlap contig gaps (36, 84, 100, 112). However,

many of these tools rely on taxon-specific metrics that are better suited for those organisms that are well studied, and therefore they lack the same resolution in identification of marker genes for organisms that are more obscure (84, 126). Other methods address potential contamination by aligning the assemblies to many references with the ability to report chimeric contigs (75). A set of standards has been proposed, called the minimum information about a single metagenomic-assembled genome (MIMAG) (14), emphasizing manual curation and review. With more studies published obtaining single-cell genomes and/or cultured representatives, we urge a systematic comparison of curated genomes with those obtained through metagenomic assembly.

Metagenomic genome assembly may still miss key aspects of the true underlying genomic variation. In order to detect low-abundance organisms, deeper sequencing is required. Shallow metagenomic sequencing, to as low as 500,000 reads, is a current alternative to amplicon sequencing in large cohort studies to gather species-level taxonomic and functional information on a large scale, at roughly the same cost as 16S rRNA sequencing (50). However, this method does not account for rare organisms and strains. Coassembly of organisms present across genomes (66), or binning of reads from many samples prior to assembly (28), has a better chance of assembling low-abundance organisms. Novel methods that use co-barcoded sequencing reads derived from individual long DNA sequences to provide the origins of reads with which to construct scaffolds (9), as well as methods that combine high-fidelity short-read sequencing with long-read sequencing data (8), will undoubtedly aid metagenomic assembly. In fact, these technologies may replace traditional whole shotgun metagenomic sequencing one day, as their costs are reduced. Obtaining information on the genomic structure of organisms with multiple chromosomes or plasmids will still require additional innovations.

Better Compositional Data Than Taxonomic Profiling Alone

Metagenomic sequencing improves the resolution of bacterial community profiling compared to 16S rRNA profiling alone and has the added advantage of being kingdom-agnostic. Despite this benefit, this leads to one of the largest challenges of the metagenomics field, classifying and quantifying the species present in a metagenomic sample. Up-to-date comparisons and benchmarking of the available tools and databases are necessary in such a fast-paced field (131).

The Diversity of Host-Associated Viruses, Fungi, and Archaea

Along with the bacterial DNA present in microbiomes, metagenomic data sets often include viral, fungal, and host DNA. The eukaryotic component has often been ignored, as genome coverage generally compares poorly to the coverage of bacterial genomes and databases containing full genomes of eukaryotes found within microbiomes are limited. Nevertheless, efforts to assemble genomes from metagenomic data sets have revealed that diverse eukaryotes can inhabit the human infant gut (80) and environments such as geothermal geysers (123). Likewise, a measurable abundance of fungi inhabiting human skin has been detected (79).

Viruses with DNA-based genomes can also be found within metagenomic data, and occasionally in high abundance. Viral genomes, mainly from bacteriophage, are small compared to bacterial or eukaryotic genomes and can be well covered. However, because they are highly diverse and fewer of the genes in bacteriophage genomes can be assigned functions or even be found in reference databases (17), their roles remain elusive. There have been efforts to gain a better understanding of this microbiome component. Numerous computational tools have been developed to identify elements of phage genomes (41, 92), either integrated prophage or free-living phage. Deeper sequencing of viral particles isolated from samples reveals the active lytic phage within a

community (16). Despite the challenges associated with analyzing phage within microbial communities, there have been several landmark observations, including a role for phage in inflammatory bowel disease pathogenesis (31). In the marine environment, metagenomic sequencing of a large number of samples, followed by co-occurrence analysis and abundance estimates of host and associated viral genomes, supports an ecological framework where lysogenic viruses dominate at high host density (29). Phage abundance data from longitudinal studies of the infant gut microbiome also resemble Lotka-Volterra dynamics (68). Assigning hosts remains a major challenge, although efforts continue to improve in this sphere (76), where even the hosts of very large, ubiquitous bacteriophage present in the human gut remained elusive until recently (30, 34, 48).

Host DNA in animal and plant systems is also sequenced along with bacterial, viral, and eukaryotic symbiont DNA. Most often, host DNA is removed prior to analysis, and this step is often required in human metagenomic studies, where host DNA can be identifiable. In certain types of microbiomes, such as the oral and skin microbiomes, this results in removal of the vast majority of sequences, up to 90% (38). There are no commonplace uses for the host DNA. Increasing numbers of studies link human genotype data with microbiome composition and/or functions to perform combined genome-wide association studies (13, 90, 120), but we have yet to observe researchers utilizing human reads from metagenomic samples to this end.

Host-Associated Phenotypes

Metagenomic data analysis has shaped our understanding of the relationship between the microbiome and host phenotypes such as health outcomes, growth, and crop productivity. While there are myriad studies ranging from medicine, agriculture, and the environment, we highlight only a few notable findings here from larger studies. The Environmental Determinants of Diabetes in the Young (TEDDY) study comprises almost 11,000 samples from 783 children beginning at 3 months of age until the onset of type 1 diabetes (T1D), or islet autoimmunity, with the goal of identifying compositional or functional aspects of the gut microbiome that may be predictive of the onset of T1D (118). It found that microbial factors associated with the onset of T1D were functionally similar but taxonomically diverse, that the gut microbiome matures faster with earlier cessation of breastfeeding, and that there are reproducible acquisitions of metabolic capabilities. Meta-analyses of smaller individual studies have also revealed important pathways to disease. Given that there have been over eight metagenomic microbiome studies on colorectal cancer on a geographically and culturally diverse set of populations, comparative metagenomic analysis can be used to find robust signals, such as an association with choline metabolism and pathways related to secondary bile acids (114, 125).

The Tara Oceans voyage was the largest metagenomic sequencing effort of marine environments to date, involving 243 size-fractionated samples that allowed for viral or prokaryotic enrichment (111). Assembly of these samples amounted to an excess of 117 million genes from over 35,000 species. Just a single drop (0.4 mL) of ocean water collected from the Sargasso Sea contained 6,236 genomes (average of 38% completeness) (82). Pairwise comparisons of the genomes found within these ocean water samples found that less than 0.1% of the genomes were from the same species, indicating the vast diversity of aquatic microbial communities. This example highlights the challenges inherent in metagenomic sequencing of microbial communities, where surveying an appropriate amount of diversity to answer a given biological question may be cost-prohibitive. Therefore, many aquatic studies on important systems like coral reefs rely on amplicon sequencing, where assessing community diversity at the species level may be more feasible (32, 47), especially in the case of time course or perturbation experiments where sampling error may complicate the interpretation of results.

Strain-Level Analyses

One of the major advantages of using metagenomic shotgun data is the ability to obtain strain-level data by resolving variations in single-nucleotide polymorphism (SNP) frequencies in microbial genomes across individuals harboring the same species. Strain-level differences can also be observed between individuals over time. There has not yet been a consensus among researchers on the most appropriate method to use, although most programs use SNPs found in single-copy core genes, either retrieved from reference genomes (1, 39, 102, 115) or taken from the sample's MAGs (18). These genes are generally phylogenetically conserved; therefore, contamination and completion are easy to determine. Many of these genes encode proteins found within the ribosome, which are rarely horizontally transferred. Within a species, the mutation frequency in these genes is often no more than one SNP per read, thus complicating phasing methods that would link sets of SNPs into single genotypes.

The study of transmission of bacterial species between environments, between hosts and the environment, or between hosts typically has relied on full genome sequences. However, headway has been made in understanding the transmission of strains within complex communities by metagenomics studies. Simple examination of the dominant strain of each organism present in a community has revealed differences in the colonization of specific species after fecal microbiota transplantation (67). Improvements to this method have shown that often people are colonized by a consortium of strains within a single species from a donor, rather than a single strain (102). Vertical transmission and colonization of strains from mother to newborn infant are observed by using SNP and strain-specific gene content methods (39, 130). These patterns have been shown to persist even in adult family members. Adult twin pairs share higher frequencies of microbial SNPs in common strains than nontwin pairs; and shared SNPs and strain-specific flexible gene content are more commonly found for species in oral and gut microbiomes of family members (128). Interestingly, metagenomic data mining has also revealed evidence of transmission between spouses, showing the malleability of the adult human microbiome (18).

Functional Assessments of Metagenomes

The main advantage of metagenomic sequencing over amplicon sequencing is the ability to perform functional profiling of microbial communities. This normally entails aligning reads to either known or de novo–assembled genes to obtain gene abundances and infer functional abundances (by merging gene abundances by gene family or function), regardless of bacterial host. In other words, unlike taxonomic profiling, these methods do not rely on marker gene sets or even assembly in some cases. Many packages exist to streamline this process (42, 52). Caution needs to be taken when performing functional profiling because up to 50% of genes within host-associated and environmental microbiomes lack annotated functions (54). For example, the earliest attempts at functionally profiling human gut microbiomes as part of the Human Microbiome Project, a large-scale effort to characterize the microbiomes of 300 individuals across several body sites with 16S rRNA and metagenomic whole-genome shotgun sequencing, led to the conclusion that functional profiles are conserved across body sites, despite vast differences in microbial composition (51). This conclusion is largely based on the portion of genes that were capable of being functionally annotated at that time, which are largely conserved core genes present in all microbes. These conclusions have largely been revised by analyzing differences between the core and distinguishing functions between body sites, and by acknowledging the extent to which genes are annotated functionally (70).

Time courses are especially useful to examine relevant changes in the microbiome that occur alongside host physiology. Metagenomic sequencing of oral, vaginal, and gut microbiome samples

of pregnant mothers reveals the dynamic nature of the microbiome during pregnancy. Aside from large intraindividual differences, gestational age of the fetus and health complications of the mother correlate with gene abundances of the mother's microbiome (45). Although the functional pathways of the various microbiota remained stable over the length of gestation, there were a few interesting examples of functions changing over time. For example, an increase in fermenter activity in the guts of the subjects over gestational time seems to suggest that fermenters could be enriched during pregnancy.

For environmental samples, a similar approach can yield insight into the functional roles of specific microbial communities and the effects of anthropogenic change on these communities. For example, metagenomic sequencing has allowed us to see that sustained warming in grasslands leads to a shift in microbial metabolic processes such as organic matter decomposition (81). Antibiotic use also has a significant effect on environmental communities, in addition to host-associated communities. Built environment microbial communities, such as those found in urban sewage systems, could be a major route for the spread of antibiotic resistance genes. Within an urban Chinese sewage system, seasonal differences of 381 different antibiotic resistance genes were found using metagenomic data, and the majority of these genes were associated with known human gut commensal bacteria (109).

Interesting Miscellanea

The microbiome has served as a unique platform for bioprospecting, and this has been aided by metagenomics approaches. In short, most analytical methods rely on examining metagenomic sequences for new enzymes that share some homology or genetic architecture with known proteins or operons. Biosynthetic gene clusters (BGCs) can be identified by observing canonical operon structures harboring sequential enzymatic processes (10). Recent advances make direct use of metagenomics reads to find potentially interesting BGCs, such as those that produce type II polyketides, which comprise clinically important drugs such as doxorubicin and tetracycline (110). Similarly, novel CRISPR-Cas systems can be identified across diverse microbiomes, enabling new functionalities and the identification of novel PAM (protospacer adjacent motif) sites that differ from the canonical NGG sequence or with more compact genetic architecture (22). Bioprospecting for biofuel enzymes across environmental metagenomic data sets can also be used to prioritize which environmental samples to use for testing (27). These types of studies will often require cloning and expressing these genes exogenously to confirm function.

Experimental methods to determine gene functionality are well defined. One example is using metagenomic data to determine target gene(s), extracting the DNA sequence, and using expression vectors to insert and express the gene in bacteria for functional screening (15, 105). Although this method has had some success with larger gene operons, including identifying certain antiproliferative, anticancer, and antibiotic compounds (26), this approach is generally more suited for those functions encoded by small operons of few genes. Samples from the environment such as soil microbiomes have been used to find new biologically and environmentally important phosphatases, as well as new domains encoding phosphatase activity, thus extending the classic categorization of known phosphatases (25).

Horizontal Gene Transfer

Horizontal gene transfer represents one of the major challenges to metagenomic assembly, yet metagenomics has been a useful tool in understanding this process. The flexible portion of a bacterium's genome allows the organism to rapidly adapt to changing environmental conditions by acquiring and incorporating novel functions, potentially altering its relationship with its host or

providing a competitive edge against other organisms, and it is therefore of high importance to microbiome researchers. Although significant progress has been made using reference genomes (3, 103) or single-cell genomes (19, 64), current methods fall short on reliably assembling mobile genetic elements and assigning mobile genetic elements to a host genome. There is great variability in mobile genetic element structure. For example, integrated transposons and certain phage comprise inverted or direct repeats and can vary between hundreds to tens of thousands of base pairs; plasmids and phage may contain large amounts of host genome. Recent evidence based on long-read metagenomic data reveals high mobility of transposable elements within a single organism, resulting in large heterogeneity within a single species in one microbiome (135). Several methods have been developed to try to apply alignment-based approaches to identify mobile genetic elements, either by examining the variation in reads aligning to reference genomes to identify flexible portions of genome assemblies (18, 33) or by aligning to reference genomes to identify those genomic regions that are not vertically conserved (106). Long-read metagenomic sequencing of microbiome samples will enable the capture of integrated mobile genetic elements and allow researchers to explore the heterogeneity of these elements within and across genomes. Additional tools, such as Hi-C sequencing, in which genomic DNA may be cross-linked and ligated with plasmid DNA (108, 129), may serve to enable better metagenomic assemblies that link plasmids with their hosts.

Replication Rates of Organisms

Relative rates of replication can be obtained using shotgun metagenomic data. Bacteria replicate their genomes bidirectionally from a singular origin of replication. Therefore, a replicating population of cells should have an abundance of metagenomic reads that map near the origin, relative to the replication terminus. This works well in cultured bacteria, and there has been success using metagenomic data, most notably in identifying replication differences across inflammatory bowel disease (IBD) and type 2 diabetes cohorts (60). Although replication rates are most readily estimated when there are phylogenetically close reference genomes with well-known replication origins, these methods have also been applied to assembled metagenomes (20). In these cases, assembled contigs need to be binned into draft genomes and then ordered according to their relative coverage to determine the overall rate of replication. There are some inherent challenges with using such a technique on MAGs that arise from incorrect binning of contigs or scaffolds and the presence of promiscuous mobile genetic elements, which may skew coverage and overestimate replication rates. The success of this approach largely depends on the quality of the MAGs, and this method is much easier to perform in microbiomes for which there are good reference genomes.

WHAT HAS METAGENOMICS MISSED?

Analyzing metagenomic data requires careful consideration of the treatment of genome assemblies and abundances. Composite MAGs can lead to inaccurate interpretations from inflated abundance or prevalence estimations, deflated diversity from ignoring or missing strain-level information, and reduced refinement in binning. Reporting quality metrics such as those proposed by the Genomics Standards Consortium (14) may lessen the burden on the end user to either determine the quality of publicly available MAGs or make incorrect assumptions. Metagenomic data sets almost exclusively rely on compositional quantification, further complicating analysis. We have only recently started reckoning with methods for assessing absolute microbial abundances and transferring this knowledge to metagenomic data (117). Aside from these technical issues, there are many aspects of

microbial ecosystems that are missed when shotgun sequencing is performed alone on microbial communities.

Ecoevolutionary Modeling

Metagenomic approaches have allowed us to obtain higher resolution than taxonomic profiling, and these approaches are poised to shed light on other aspects of microbial community assembly and evolutionary trajectories, yet the approaches are still fairly nascent in this regard. A remaining challenge is that information derived from individual genomes, which can be crucial for ecological or evolutionary inferences, can be lost. For example, population variation in genetic architecture, mobile genetic elements, and SNP diversity may be difficult to ascertain. There have been several early efforts to draw evolutionary models from metagenomic data alone (44). From an examination of strain-level differences across gut microbiome samples, it appears that gene gains and losses are fairly common and can sweep to high frequencies relatively quickly, though strain replacement is the more dominant trend over longer periods of time. Differences in gene copy number variants within microbial genomes have proved to be informative about the function, and possibly the evolutionary trajectories, of specific organisms (46, 132). To some extent, it will take time for ecoevolutionary theory to develop, as we are still learning about the genomic structure in microbial communities in their natural environment. For example, a large number of small genes were recently uncovered in metagenomic data sets, many of whose functions are unknown (94).

Phenotype and Comprehensive Functional Profiling

Phenotype is a complex trait, and metagenomic data alone are often insufficient in determining phenotypic traits. As an example, many of the metabolites in the human gut microbiome have strong associations with microbial species and pathways present in metagenomic data (119), but predictions of metabolic output using metagenomic data alone can have high variance (73). First, phenotypic differences may be driven by large differences at the level of transcription. For example, *Verrucomicrobia* was identified as highly abundant in soil communities, leading to the assumption that it was vital for soil health and functioning. However, metatranscriptomics analysis revealed that *Verrucomicrobia* is metabolically inactive in the soil. As another example, metatranscriptomics of ruminant livestock showed that a mix of bacterial, archaeal, and eukaryotic species are active during plant degradation and methane production, which may be missed when focusing on either bacteria or eukaryotes alone (104). Second, gene-gene interactions across species may drive specific outputs of pathways, yet few of these in natural microbial communities are known. Examining co-occurrence networks may provide clues to codependent organisms (37, 43, 63), yet these methods have not been applied widely to metagenomic data sets. Modeling metabolic outputs using metagenomic data (71) is another important step toward this end, yet these models tend to be more accurate for less diverse microbiomes, such as those found in termites (61).

Despite these considerations, the predominant limitation in translating metagenomic data to phenotype is the overall proportion of genes we can annotate. KEGG (56), COG (113), PFAM (40), TIGRFAM (49), MetaCyc (58), and other databases used to assign functions to assembled genes only capture roughly half of functions in a commonly assayed microbiome, such as the human gut (54); however, they capture a much smaller fraction in diverse, less-sampled microbiomes such as those from certain soil communities, less-studied animal microbiomes, and those from human populations living in low- and middle-income countries or remote areas (77, 85). Large-scale functional studies will be vital to improving functional databases, but these experiments are laborious, and curation of these databases is often done manually.

Alternative methods have been applied to microbial communities to gain additional functional insight. Stable isotope probing using isotopically labeled substrates can inform researchers about the specific bacteria utilizing the substrate (87, 124). The labeled substrate gets incorporated into DNA that can be separated and sequenced. One interesting example is the identification of new bile salt hydrolase genes in the gut microbiome using probes that label active enzymes that can be assayed with proteomic tools and metagenomic sequencing to identify those proteins (83). To determine which organisms are metabolically active in a sample, PMA (propidium monoazide) has been used to distinguish between live and dead cells. It intercalates DNA, but only in those cells with compromised membranes. Light exposure causes covalent bonds to form with the DNA, resulting in fragmentation and rendering it unamenable to DNA sequencing. This method has been used widely for samples prior to 16S rRNA amplicon sequencing, but it was recently used to identify the live portion of a saliva microbiome sample that underwent metagenomic sequencing (74). Examples that probe specific functions will enhance our understanding of metagenomic communities above metagenomics alone.

Similarly, integrating metagenomic, metatranscriptomic, and metabolomic data can alternatively improve functional assessments of communities. There has been a sharp increase of methods that integrate omics data. The multi-omics approach is valuable in that it does not require bacterial culturing, which is an impediment to examining microbiome function. In phase 2 of the Human Microbiome Project, known as the Integrative Human Microbiome Project (iHMP), 1,785 individuals from three microbiome-associated condition cohorts were sampled: pregnancy and preterm birth, IBD, and type 2 diabetes. A wealth of data was collected, including gut microbiome metagenomes, metatranscriptomes, proteomes, metabolomes, and virome data (69). By integrating these data, the authors were able to associate functions and molecular dynamics to specific taxa of the gut microbiome. Similarly, omics studies reveal that twin pairs share metabolic pathways on average almost twice as much as they share species (119), suggesting that in the search for therapeutic targets, genetic associations, or biomarkers, it may be more informative to study the functions of the gut microbiome rather than the organismal composition or species diversity.

In addition to omics performed on microbial communities, an increased number of studies are also incorporating measurements of the host. Zhou et al. (134) used a longitudinal multi-omics approach to study host-microbe dynamics in prediabetes. By integrating metagenomes, transcriptomes, metabolomes, cytokines, and proteomes from 106 individuals, they were able to detect molecular signatures in 1 person that preceded the onset of type 2 diabetes, which included the inflammation markers interleukin-1 receptor agonist and high-sensitivity C-reactive protein. More broadly, they were able to characterize thousands of host-microbe interactions that were distinctive between insulin-sensitive and insulin-resistant individuals. Techniques to integrate the diversity of data sources, each with their own benefits and limitations, are still under development (7).

Spatial Analyses of the Microbiome

Missing from many metagenomic analyses are temporal and spatial dynamics. Time course experiments are relatively expensive, but several large-scale temporal data sets are starting to emerge, such as a recent study of patients with IBD (89). Similar to 16S rRNA amplicon data sets, metagenomic time course data analysis requires not only careful managing of the compositionality but also autocorrelation. Techniques to obtain spatial data about microbiomes are emerging (91, 97, 98, 121, 122), by applying species-level probes or by sequencing proximate microbes captured in preserved microscale blocks or by performing laser dissection of fixed communities. These techniques capture species-level interactions and have not yet scaled to accommodate metagenomic sequencing approaches.

INNOVATION AND FUTURE DIRECTIONS

Metagenomic analyses of microbiomes will be vastly altered in the coming decade by technological and accessibility improvements in DNA and RNA sequencing. As long-read sequencing becomes cheaper, it will negate the need for elaborate methods for genome assembly and phasing of SNPs. Since the quality of draft genomes assembled from short-read sequencing data may be highly variable, and few studies incorporate reference genomes for comparison, long-read sequencing will go a long way to improve the quality of these metagenomic assembled genomes. The increased use of metagenomics will create new challenges in data storage and data reporting, especially as the types of data platforms (e.g., short-read, long-read, Hi-C, Tn-seq, functional screening) grow. The size of future data sets will necessitate solutions for data compression, high-speed search, and memory-efficient assembly methods, some of which are starting to become available (6, 28, 59). Standard protocols for data reporting and submission will be important, especially in terms of what information and metadata to provide.

If the expenses and error rates associated with long-read sequencing are reduced, many of the challenges associated with short-read sequencing will fade. Assemblies will be less fragmented, SNP phasing will be inferred based on co-occurrence on reads, and integrated mobile genetic elements will be associated with their flanking genomes. Nevertheless, this will not fully solve the problem of associating extrachromosomal elements with their host genomes. Alternatively, single-cell genome sequencing may provide the technological advance that surmounts some of these problems. Yet, even the largest of studies is several thousand cells, orders of magnitude below what is typically sampled in a shotgun metagenomic sample. Currently single-cell sequencing technologies are limited by the cost, and the quality of the genomes is highly variable, resulting in a large amount of data loss.

Despite these projected improvements, methods to assign functions to the vast number of genes within the microbiome are still necessary to understand the mechanisms underlying microbiome-related phenotypic outcomes. It was recently discovered that a single-amino-acid-residue difference in the dopamine dehydroxylase (DahD) gene in *Eggerthella lenta* altered whether the pharmaceutical L-dopa remained active in microbiome samples from a cohort of patients with Parkinson disease (72). This level of detailed understanding of gene function will be required for the research field to go beyond characterization of microbiomes to understanding the mechanisms underlying an overall phenotype. Examination of modifications of DNA, such as methylation patterns obtained using single-molecule real-time (SMRT) sequencing (5), can reveal interesting patterns of plasmid mobility within natural microbial communities. Technological innovation will reveal interesting layers of organismal interactions, functional roles, evolutionary trajectories, and niche occupancy in microbial communities, which will lead to a better understanding of how to shape communities to achieve a prosperous outcome for health, agriculture, or environmental sustainability.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

F.N.N. is a Cornell Graduate Dean's Scholar. I.L.B. is an Atkinson Center for a Sustainable Future Faculty Fellow, a Pew Scholar in the Biomedical Sciences, a Packard Fellow for Science and Engineering, an NIH New Innovator, and a Sloan Foundation Fellow. We thank members of the Brito lab for their thoughtful comments on our manuscript.

LITERATURE CITED

1. Albanese D, Donati C. 2017. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* 8(1):2260
2. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, et al. 2014. Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11(11):1144–46
3. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. 2019. A reverse ecology approach based on a biological definition of microbial populations. *Cell* 178(4):820–34.e14
4. Ayling M, Clark MD, Leggett RM. 2019. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* 21(2)584–94
5. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang X-S, et al. 2018. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* 36(1):61–69
6. Berger B, Peng J, Singh M. 2013. Computational solutions for omics data. *Nat. Rev. Genet.* 14(5):333–46
7. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, et al. 2016. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinform.* 17(Suppl. 2):15
8. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, et al. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 37(8):937–44
9. Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, et al. 2018. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. Biotechnol.* 36(11):1067–80
10. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, et al. 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* 47(W1):W81–87
11. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13(12):R122
12. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37(8):852–57
13. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, et al. 2016. The effect of host genetics on the gut microbiome. *Nat. Genet.* 48(11):1407–12
14. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35(8):725–31
15. Brady SF. 2007. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* 2(5):1297–305
16. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, et al. 2003. Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185(20):6220–23
17. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. 2002. Genomic analysis of uncultured marine viral communities. *PNAS* 99(22):14250–55
18. Brito IL, Gurry T, Zhao S, Huang K, Young SK, et al. 2019. Transmission of human-associated microbiota along family and social networks. *Nat. Microbiol.* 4(6):964–71
19. Brito IL, Yilmaz S, Huang K, Xu L, Jupiter SD, et al. 2016. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535(7612):435–39
20. Brown CT, Olm MR, Thomas BC, Banfield JF. 2016. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* 34(12):1256–63
21. Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, et al. 2016. Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* 533(7604):543–46
22. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, et al. 2017. New CRISPR-Cas systems from uncultivated microbes. *Nature* 542(7640):237–41
23. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11(12):2639–43
24. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13(7):581–83

25. Castillo Villamizar GA, Nacke H, Boehning M, Herz K, Daniel R. 2019. Functional metagenomics reveals an overlooked diversity and novel features of soil-derived bacterial phosphatases and phytases. *mBio* 10(1):e01966-18
26. Chang F-Y, Brady SF. 2013. Discovery of indolotryptoline antiproliferative agents by homology-guided metagenomic screening. *PNAS* 110(7):2478
27. Chaudhary N, Gupta A, Gupta S, Sharma VK. 2017. BioFuelDB: a database and prediction server of enzymes involved in biofuels production. *PeerJ*. 5:e3497
28. Cleary B, Brito IL, Huang K, Gevers D, Shea T, et al. 2015. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* 33(10):1053–60
29. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, et al. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun*. 8(1):15955
30. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, et al. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 4(4):693–700
31. Duerkop BA, Kleiner M, Paez-Espino D, Zhu W, Bushnell B, et al. 2018. Murine colitis reveals a disease-associated bacteriophage community. *Nat. Microbiol.* 3(9):1023–31
32. Dunphy CM, Gouhier TC, Chu ND, Vollmer SV. 2019. Structure and stability of the coral microbiome in space and time. *Sci. Rep.* 9(1):6785
33. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* 27(1):140–53.e9
34. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, et al. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5(1):4498
35. Edgar RC. 2016. UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv 074252
36. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, et al. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 3:e1319
37. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, et al. 2012. Microbial co-occurrence relationships in the human microbiome. *PLOS Comput. Biol.* 8(7):e1002606
38. Ferretti P, Farina S, Cristofolini M, Girolomoni G, Tett A, Segata N. 2017. Experimental metagenomics and ribosomal profiling of the human skin microbiome. *Exp. Dermatol.* 26(3):211–19
39. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, et al. 2018. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* 24(1):133–145.e5
40. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42(D1):D222–30
41. Fouts DE. 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*. 34(20):5839–51
42. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, et al. 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15(11):962–68
43. Friedman J, Alm EJ. 2012. Inferring correlation networks from genomic survey data. *PLOS Comput. Biol.* 8(9):e1002687
44. Garud NR, Good BH, Hallatschek O, Pollard KS. 2019. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLOS Biol*. 17(1):e3000102
45. Goltsman DSA, Sun CL, Proctor DM, DiGiulio DB, Robaczewska A, et al. 2018. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res*. 28(10):1467–80
46. Greenblum S, Carr R, Borenstein E. 2015. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 160(4):583–94
47. Grottoli AG, Dalcin Martins P, Wilkins MJ, Johnston MD, Warner ME, et al. 2018. Coral physiology and microbiome dynamics under combined warming and ocean acidification. *PLOS ONE* 13(1):e0191156
48. Guerin E, Shkoporov A, Stockdale SR, Gonzalez-Tortuero E, Ross RP, Hill C. 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. *Cell Host Microbe* 24:653–64

49. Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31(1):371–73
50. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, et al. 2018. Evaluating the information content of shallow shotgun metagenomics. *mSystems* 3(6):e00069-18
51. Hum. Microbiome Proj. Consort., Huttenhower C, Gevers D, Knight R, Abubucker S, et al. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* 486(7402):207–14
52. Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17(3):377–86
53. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ.* 2:e603
54. Joice R, Yasuda K, Shafquat A, Morgan XC, Huttenhower C. 2014. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* 20(5):731–41
55. Jump. Consort. Hum. Microbiome Proj. Data Gener. Work. Group. 2012. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLOS ONE* 7(6):e39315
56. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44(D1):D457–62
57. Kang DD, Li F, Kirton E, Thomas A, Egan R, et al. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359
58. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. 2002. The MetaCyc database. *Nucleic Acids Res.* 30(1):59–61
59. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26(12):1721–29
60. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, et al. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* 349(6252):1101–6
61. Kundu P, Manna B, Majumder S, Ghosh A. 2019. Species-wide metabolic interaction network for understanding natural lignocellulose digestion in termite gut microbiota. *Sci. Rep.* 9(1):16329
62. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, et al. 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.* 14(4):169–81
63. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.* 11(5):e1004226
64. Labonté JM, Field EK, Lau M, Chivian D, Van Heerden E, et al. 2015. Single cell genomics indicates horizontal gene transfer and viral infections in a deep subsurface Firmicutes population. *Front. Microbiol.* 6:349
65. Lagier J-C, Dubourg G, Million M, Cadoret F, Bilen M, et al. 2018. Culturing the human microbiota and culturomics. *Nat. Rev. Microbiol.* 16(9):540–50
66. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10):1674–76
67. Li SS, Zhu A, Benes V, Costea PI, Hercog R, et al. 2016. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* 352(6285):586–89
68. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, et al. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* 21(10):1228–34
69. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, et al. 2019. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569(7758):655–62
70. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, et al. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* 550(7674):61–66
71. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, et al. 2017. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* 35(1):81–89
72. Maini Rekdal V, Bess EN, Bisanz JE, Turnbaugh PJ, Balskus EP. 2019. Discovery and inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. *Science* 364(6445):eaau6323
73. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, et al. 2019. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10(1):3136

74. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. 2018. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* 6(1):42
75. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32(7):1088–90
76. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. *PLOS Genet.* 9(12):e1003987
77. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. 2019. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568(7753):505–10
78. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27(5):824–34
79. Oh J, Byrd AL, Deming C, Conlan S, NISC Comp. Seq. Program, et al. 2014. Biogeography and individuality shape function in the human skin metagenome. *Nature* 514(7520):59–64
80. Olm MR, West PT, Brooks B, Firek BA, Baker R, et al. 2019. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* 7(1):26
81. Orellana LH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT. 2018. Year-round shotgun metagenomes reveal stable microbial communities in agricultural soils and novel ammonia oxidizers responding to fertilization. *Appl. Environ. Microbiol.* 84(2):e01646-17
82. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, et al. 2019. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 179(7):1623–35.e11
83. Parasar B, Zhou H, Xiao X, Shi Q, Brito IL, Chang PV. 2019. Chemoproteomic profiling of gut microbiota-associated bile salt hydrolase activity. *ACS Cent. Sci.* 5(5):acscentsci.9b00147
84. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25(7):1043–55
85. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, et al. 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176(3):649–62.e20
86. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28(11):1420–28
87. Pepe-Ranney C, Campbell AN, Koechli CN, Berthrong S, Buckley DH. 2016. Unearthing the ecology of soil microorganisms using a high resolution DNA-SIP approach to explore cellulose and xylose metabolism in soil. *Front. Microbiol.* 7:703
88. Porras AM, Brito IL. 2019. The internationalization of human microbiome research. *Curr. Opin. Microbiol.* 50:50–55
89. Poyet M, Groussin M, Gibbons SM, Avila-Pacheco J, Jiang X, et al. 2019. A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* 25(9):1442–52
90. Qin J, Li Y, Cai Z, Li S, Zhu J, et al. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490(7418):55–60
91. Riva A, Kuzyk O, Forsberg E, Siuzdak G, Pfann C, et al. 2019. A fiber-deprived diet disturbs the fine-scale spatial architecture of the murine colon microbiome. *Nat. Commun.* 10(1):4366
92. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 3:e985
93. Sarhan MS, Hamza MA, Youssef HH, Patz S, Becker M, et al. 2019. Culturomics of the plant prokaryotic microbiome and the dawn of plant-based culture media—a review. *J. Adv. Res.* 19:15–27
94. Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, et al. 2019. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* 178(5):1245–59.e14
95. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75(23):7537–41
96. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, et al. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat. Methods* 14(11):1063–71

97. Sheth RU, Li M, Jiang W, Sims PA, Leong KW, Wang HH. 2019. Spatial metagenomic characterization of microbial biogeography in the gut. *Nat. Biotechnol.* 37(8):877–83
98. Shi H, Zipfel W, Brito I, Vlaminck I De. 2019. Highly multiplexed spatial mapping of microbial communities. bioRxiv 678672
99. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, et al. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3(7):836–43
100. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–12
101. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19(6):1117–23
102. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, et al. 2018. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 23(2):229–40.e5
103. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* 480(7376):241–44
104. Söllinger A, Tveit AT, Poulsen M, Noel SJ, Bengtsson M, et al. 2018. Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. *mSystems* 3(4):e00038-18
105. Sommer MOA, Dantas G, Church GM. 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 325(5944):1128–31
106. Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiomie* 7(1):36
107. Song W-Z, Thomas T. 2017. Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* 33(12):1873–75
108. Stalder T, Press MO, Sullivan S, Liachko I, Top EM. 2019. Linking the resistome and plasmidome to the microbiome. *ISME J.* 13(10):2437–46
109. Su J-Q, An X-L, Li B, Chen Q-L, Gillings MR, et al. 2017. Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome* 5(1):84
110. Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, et al. 2019. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. *Science* 366(6471):eaax9176
111. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, et al. 2015. Structure and function of the global ocean microbiome. *Science* 348(6237):1261359
112. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, et al. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10(12):1196–99
113. Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28(1):33–36
114. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, et al. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* 25(4):667–78
115. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27(4):626–38
116. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6(1):158
117. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, et al. 2017. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* 551(7681):507–11
118. Vatanen T, Franzosa EA, Schwager R, Tripathi S, Arthur TD, et al. 2018. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* 562(7728):589–94
119. Visconti A, Le Roy CI, Rosa F, Rossi N, Martin TC, et al. 2019. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* 10(1):4505

120. Weissbrod O, Rothschild D, Barkan E, Segal E. 2018. Host genetics and microbiome associations through the lens of genome wide association studies. *Curr. Opin. Microbiol.* 44:9–19
121. Welch JLM, Hasegawa Y, McNulty NP, Gordon JI, Borisy GG. 2017. Spatial organization of a model 15-member human gut microbiota established in gnotobiotic mice. *PNAS* 114(43):E9105–14
122. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. 2016. Biogeography of a human oral microbiome at the micron scale. *PNAS* 113(6):E791–800
123. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. 2018. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* 28(4):569–80
124. Wilhelm RC, Singh R, Eltis LD, Mohn WW. 2019. Bacterial contributions to delignification and lignocellulose degradation in forest soils with metagenomic and quantitative stable isotope probing. *ISME J.* 13(2):413–29
125. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, et al. 2019. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* 25(4):679–89
126. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20(1):257
127. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32(4):605–7
128. Xie H, Guo R, Zhong H, Feng Q, Lan Z, et al. 2016. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* 3(6):572–84.e3
129. Yaffe E, Relman DA. 2019. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. *Nat. Microbiol.* 5(2):343–53
130. Yassour M, Jason E, Hogstrom LJ, Arthur TD, Tripathi S, et al. 2018. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe* 24(1):146–54.e4
131. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking metagenomics tools for taxonomic classification. *Cell* 178(4):779–94
132. Zeevi D, Korem T, Godneva A, Bar N, Kurilshikov A, et al. 2019. Structural variation in the gut microbiome associates with host health. *Nature* 568(7750):43–48
133. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–29
134. Zhou W, Sailani MR, Contrepois K, Zhou Y, Ahadi S, et al. 2019. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* 569(7758):663–71
135. Zlitni S, Bishara A, Moss EL, Tkachenko E, Kang JB, et al. 2020. Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Med.* 12:50