

REVIEW ARTICLES

High-resolution characterization of the human microbiome



CECILIA NOECKER, COLIN P. MCNALLY, ALEXANDER ENG, and ELHANAN BORENSTEIN

SEATTLE, WA AND SANTA FE, NM

The human microbiome plays an important and increasingly recognized role in human health. Studies of the microbiome typically use targeted sequencing of the 16S rRNA gene, whole metagenome shotgun sequencing, or other meta-omic technologies to characterize the microbiome's composition, activity, and dynamics. Processing, analyzing, and interpreting these data involve numerous computational tools that aim to filter, cluster, annotate, and quantify the obtained data and ultimately provide an accurate and interpretable profile of the microbiome's taxonomy, functional capacity, and behavior. These tools, however, are often limited in resolution and accuracy and may fail to capture many biologically and clinically relevant microbiome features, such as strain-level variation or nuanced functional response to perturbation. Over the past few years, extensive efforts have been invested toward addressing these challenges and developing novel computational methods for accurate and high-resolution characterization of microbiome data. These methods aim to quantify strain-level composition and variation, detect and characterize rare microbiome species, link specific genes to individual taxa, and more accurately characterize the functional capacity and dynamics of the microbiome. These methods and the ability to produce detailed and precise microbiome information are clearly essential for informing microbiome-based personalized therapies. In this review, we survey these methods, highlighting the challenges each method sets out to address and briefly describing methodological approaches. (Translational Research 2017;179:7–23)

Abbreviations: CNV = copy number variation; FISH = fluorescent in situ hybridization; HMM = hidden Markov model; KEGG = Kyoto Encyclopedia of Genes and Genomes; LCA = lowest common ancestor; OTU = operational taxonomic unit; rRNA = ribosomal RNA; SNP = single-nucleotide polymorphism

From the Department of Genome Sciences, University of Washington, Seattle, WA; Department of Computer Science and Engineering, University of Washington, Seattle, WA; Santa Fe Institute, Santa Fe, NM.

Cecilia Noecker, Colin P. McNally, and Alexander Eng contributed equally to this work.

Submitted for publication April 29, 2016; revision submitted July 12, 2016; accepted for publication July 15, 2016.

Reprint requests: Elhanan Borenstein, Genome Sciences, University of Washington, Seattle, WA 98195-5065; e-mail: elbo@uw.edu.

1931-5244/\$ - see front matter

© 2016 Elsevier Inc. All rights reserved.

<http://dx.doi.org/10.1016/j.trsl.2016.07.012>

INTRODUCTION

Recent marked advances in sequencing technologies have been followed by an explosion of studies using these technologies to explore a wide range of microbial communities, including those that inhabit the human body. Such studies apply targeted sequencing of the 16S rRNA gene and whole metagenome shotgun sequencing to characterize the human microbiome in numerous settings. Analyses of these sequencing data commonly use an assortment of clustering, binning, annotation, and assembly algorithms to ultimately profile the composition of species in each sample, the set of genes they collectively encode,

or the genome sequence of specific member species (Fig 1). Taken together, these efforts to map the human microbiome in health and in disease have led to an increased appreciation for the important role of the microbiome in human well-being.¹⁻⁵

Nevertheless, common computational metagenomic analysis methods are often limited in resolution and may fail to resolve nuanced, yet important and potentially clinically relevant details concerning the composition of species and genes in the microbiome. Standard 16S rRNA surveys, for example, are often limited to a genus-level taxonomic identification,⁶ can fail to distinguish closely related taxonomic groups, and cannot always unambiguously discriminate rare, low-abundance taxa from noise.⁷ Shotgun metagenomic analyses may similarly fail to identify the taxonomic origins of a gene of interest or to produce accurate and unbiased estimates of gene families' abundances.^{8,9}

Clearly, however, given the complexity of the human microbiome, accurate and high-resolution mapping of the microbiome is crucial for gaining a principled understanding of community behavior, function, and ultimately its impact on the host.¹⁰ For example, accurately profiling strain-level microbiome composition is vital for tracking ecological trends over time, such as the spread of bacterial vaginosis-associated strains between sexual partners.¹¹ Discerning subtle genomic variation between closely related strains of the same species may also have important clinical implications,

as in the case of *Propionibacterium acnes*, which displays extensive strain variation in the skin microbiome with potential impact on various skin conditions.¹² Likewise, *Escherichia coli* has well-characterized variation in toxin production, which results in high pathogenicity for a subset of strains, whereas other strains are commonly found in healthy gut microbiomes.¹³ Careful differentiation of strains can also inform clinical decision making by, for example, providing valuable insights as to whether a patient will respond to the heart failure drug digoxin.¹⁴ Accurate detection of low-abundance species is similarly essential as such rare species may still play important roles in various biological processes. Indeed, even species present at less than 0.01% abundance in oral microbial communities can play a key role in causing oral inflammatory disease.¹⁵ A high-quality, unbiased, and rigorous characterization of the metagenome's gene content is equally important for pinpointing disease-associated shifts in the functional capacity of the microbiome.⁹

Moreover, many molecular processes that play important roles in the microbiome's activity and dynamics go beyond the microbiome's taxonomic and genic composition and accordingly cannot be profiled through metagenome sequencing. For example, oligosaccharides found in breast milk can change microbial gene expression and production of physiologically relevant microbial metabolites in the infant gut without affecting the abundance of most species.¹⁶ Exploring such processes

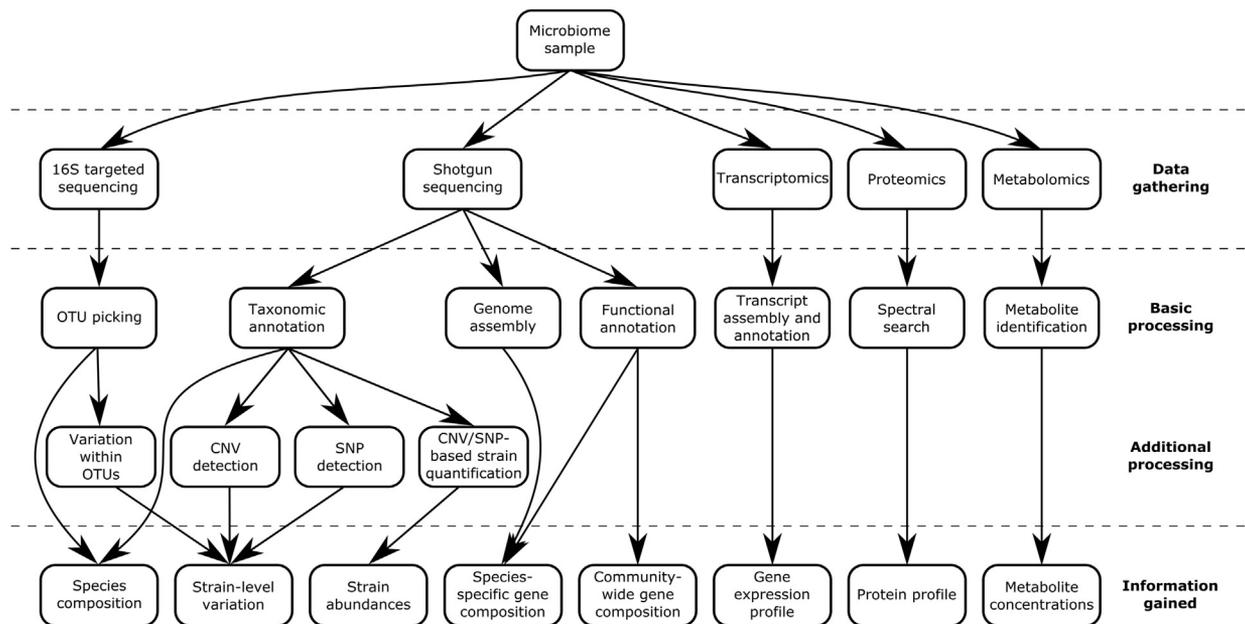


Fig 1. Schemes of microbiome analysis. Metagenomic data and other meta-omic data can be processed and analyzed in various ways to address a diverse set of questions concerning the microbiome's composition, capacity, and function.

Table 1. Computational tools for characterizing the human microbiome

Tool	Synopsis	Reference	Website
16S rRNA analysis			
M-pick	Identifies taxonomic clusters based on modularity analysis of a sequence read graph	17	http://plaza.ufl.edu/xywang/Mpick.htm
Swarm	Uses iterative linkage clustering to identify natural subpopulations	18	https://github.com/torognes/swarm
Minimum entropy decomposition	Hierarchically clusters sequences by iteratively subdividing based on sequence entropy	19	http://merenlab.org/2014/11/04/med/
Oligotyping	Identifies subpopulations within predefined OTUs by identifying and clustering the most informative nucleotide positions	20	http://merenlab.org/projects/oligotyping/
Oclust	Hierarchically clusters PacBio circular consensus sequencing reads into OTUs	21	https://github.com/oscar-franzen/oclust/
CopyRighter	Correct 16S rRNA data for CNV	22	https://github.com/fangly/AmpliCopyrighter
metagenomeSeq	R package for normalization and differential abundance analysis	23	http://cbcb.umd.edu/software/metagenomeSeq/
PhyloSeq	R package for processing, normalization, differential abundance analysis, and visualization	24	https://joey711.github.io/phyloseq/
RAIDA	Differential abundance analysis using ratios between taxa	25	http://cals.arizona.edu/~anling/sbg/software.htm
Strain-level characterization using shotgun metagenomic data			
WGFAST	Identified strains from low-coverage genome data sets	26	https://github.com/jasonsahl/wgfast
Sigma	Taxonomic analysis of metagenome data at the strain level, including variant calling and statistical uncertainty calculations	27	http://sigma.omicsbio.org
ConStrains	Identifies strains from metagenomic sequence data and reconstructs their phylogeny	28	https://bitbucket.org/luo-chengwei/constrains
PathoScope	Identifies the proportion of reads from individual microbial strains in metagenomic sequencing data	29	https://sourceforge.net/projects/pathoscope/
—	Large-scale characterization of strain-level CNV	30	http://elbo.gs.washington.edu/download.html
PhyloCNV	Profiles species abundance; identifies, characterizes, and analyzes strains based on single-nucleotide polymorphisms and CNV	31	https://github.com/snayfach/PhyloCNV
Assembling reference genomes from shotgun metagenomic data			
MEGAHIT	Assembles metagenomic short reads with low memory use	32	https://github.com/voutcn/megahit
MetaVelvet	Assembles metagenomic short reads into contigs	33	http://metavelvet.dna.bio.keio.ac.jp/
CONCOCT	Binning using Gaussian mixture models and sequence composition and abundance	34	https://github.com/BinPro/CONCOCT
GroopM	Automated genome recovery from metagenomes	35	http://ecogenomics.github.io/GroopM/
MetaBAT	Bins genomes based on tetranucleotide frequency and abundance	36	https://bitbucket.org/berkeleylab/metabat
MaxBin	Bins assembled genomes using EM algorithm	37	http://downloads.jbei.org/data/microbial_communities/MaxBin/MaxBin.html

(Continued)

Table 1. (Continued)

Tool	Synopsis	Reference	Website
ABAWACA	Binning based on mono, di, and tri-nucleotide frequency and abundance, using hierarchical clustering	38	https://github.com/CK7/abawaca
CheckM	Assessing the quality of putative genomes	39	http://ecogenomics.github.io/CheckM/
MetaDecon	Metagenomic deconvolution and reconstruction of species-specific genomic content	40	http://elbo.gs.washington.edu/software_metadecon.html
Specialized functional annotation HMM databases			
FOAM	Annotation of KEGG orthology groups	41	http://cbb.pnnl.gov/portal/software/FOAM.html
Resfams	Annotation of antibiotic resistance genes	42	http://www.dantaslab.org/resfams
dbCAN	Annotation of carbohydrate-active enzyme (CAZyme) genes	43	http://csbl.bmb.uga.edu/dbCAN/annotate.php
Functional annotation and quantification of shotgun metagenomic data			
ShortBRED	Representative marker-based protein family profiling	44	https://huttenhower.sph.harvard.edu/shortbred
MUSICC	Accurate normalization of functional profiles	9	http://elbo.gs.washington.edu/software_musicc.html
MicrobeCensus	Accurate normalization of functional profiles	45	https://github.com/snayfach/MicrobeCensus
Combined taxonomic and functional annotation of shotgun metagenomic data			
LMAT	<i>k</i> -mer-based taxonomic assignment of metagenomic reads	46	https://sourceforge.net/projects/lmat/
Kraken	<i>k</i> -mer-based taxonomic assignment of metagenomic reads	47	http://ccb.jhu.edu/software/kraken/
TAC-ELM	Neural network-based taxonomic classification of metagenomic reads	48	http://cs.gmu.edu/~mlbio/TAC-ELM/
MetAnnotate	Combined taxonomic and functional annotation with HMMs and alignment to HMM-family protein sequences	49	http://metannotate.uwaterloo.ca
SeMeta	Multiple-level clustering of reads followed by cluster assignment to taxa through representative read alignment	50	http://it.hcmute.edu.vn/bioinfo/metapro/SeMeta.html
MetaCluster-TA	Clusters assemble contigs and assign taxonomy by alignment	51	http://i.cs.hku.hk/~alse/MetaCluster/index.html
Meta-omic analysis tools			
TAG	Assembles a metatranscriptome incorporating information from a metagenome assembly	52	http://omics.informatics.indiana.edu/TAG/
Anvi'o	Interactive processing and visualization of metagenomes and metatranscriptomes	53	http://merenlab.org/projects/anvio/
Pipasic	Produces quantitative strain-specific peptide assignments using a sequence similarity correction	54	https://sourceforge.net/projects/pipasic/
BacSpace	Processing and quantitative analysis of microbial community FISH imaging data	55	https://bitbucket.org/kchuanglab/bacspace/downloads
MIMOSA	Metabolic model-based integration of microbiome taxonomic and metabolomic profiles	56	http://elbo.gs.washington.edu/software_MIMOSA.html

Abbreviations: CNV, copy number variation; FISH, fluorescent in situ hybridization; HMM, hidden Markov model; OTUs, operational taxonomic units; rRNA, ribosomal RNA.

may require detailed information about transcript and protein variation, metabolite concentrations, and spatial distribution. Indeed, new “omic” technologies can now comprehensively quantify such community features, but computational methods available to analyze the

resulting sizable and novel data sets are largely still in early stages and may be limited in resolution.

In this review, we accordingly describe an array of recent computational methods and analytical approaches that set out to address these challenges and

to provide high-resolution, multi-omic, systematic characterizations of the microbiome at multiple levels (Table 1). Although some of these approaches have primarily been applied to environmental microbial communities, all are broadly applicable and potentially useful in the context of the human microbiome and its health impacts. We first discuss taxonomic analysis of the microbiome, focusing on methods for detecting strain-level variation within each member species. We specifically describe methods that use targeted 16S rRNA or whole metagenome sequencing data for strain-level profiling, identification, and tracking, either *de novo* or based on existing reference genomes. We also describe recent methods for assembling the genomes of novel strains directly from metagenomic data. We next discuss methods for improved functional characterization of the microbiome, including accurate detection of the various gene families encoded by the metagenome and precise quantification of their abundances, and for linking taxonomic and functional profiles. Finally, we describe several recent frameworks for analyzing and integrating other microbiome-derived high-throughput omic data sets and for profiling additional facets of the microbiome's composition and activity.

HIGH-RESOLUTION CHARACTERIZATION OF THE MICROBIOME'S TAXONOMIC COMPOSITION

One of the most common and relatively accessible starting points for human microbiome analysis is taxonomic profiling. Specifically, by sequencing and analyzing taxonomy-associated marker genes, researchers can readily identify the various species present in a given microbiome sample and estimate the relative abundances of each species.⁵⁷ The study of such taxonomic profiles and the way they vary across individuals or between cohorts can provide numerous insights into the link between the microbiome ecology and the host's health. Such studies can, for example, pinpoint specific species with known virulence factors or community-wide dysbiotic features as biomarkers of disease.^{58,59} As noted previously, however, taxonomic profiling is often limited in resolution and may therefore hinder our ability to detect more fine-grained determinants of disease. Subsequently, we describe several new and exciting developments in the analysis of both marker gene data and whole metagenomes that aim to provide a more detailed, high-resolution map of the microbiome's taxonomy.

High-resolution and accurate analysis of 16S rRNA data. To date, the most prevalent form of comprehensive microbiome taxonomic data is produced via targeted amplification and sequencing of the 16S rRNA gene, a commonly used phylogenetic marker.⁶⁰ The

analysis of such 16S rRNA sequencing data typically involves clustering of the obtained sequences (usually based on sequence overall percent similarity) into operational taxonomic units (OTUs) and determining the relative abundance of each OTU in the sample. The taxonomy of each OTU can then be inferred by clustering reads with reference sequences of known taxonomy or by a classifier algorithm that predicts each OTU's (or each read's) likely taxonomy.^{61,62}

This clustering-based approach is efficient, widely used, and well established; yet several challenges remain in terms of accurate and precise taxonomic quantification at the species or strain level. First, a measure of the overall percent similarity between two 16S rRNA sequences may not fully capture the variation present in the sequenced region or the taxonomic divergence that this variation represents. Indeed, the number and nature of polymorphisms and of true subpopulations included within a single, similarity-based OTU cluster can vary greatly across OTUs.⁶³ A number of recently introduced algorithms aim to account for such variation using graph-based clustering approaches and grouping sequences based on local base differences between reads rather than by overall percent similarity.^{17,64} These algorithms have proved successful in identifying higher resolution sequence clusters and more accurately describing the population structure in each sample. One example of such an algorithm, termed *Swarm*,^{18,65} first performs exact linkage clustering to group reads that have one nucleotide differences to any other read in the same cluster and then refines each cluster based on read abundance distributions. This approach has been successfully applied to characterize fine-scale taxonomic profiles of bacteria and protists in several environments.^{66,67} An alternative method for addressing the limitation of similarity-based clustering, termed *minimum entropy decomposition*, generates a hierarchy of read groupings by iteratively subdividing the data set into groups based on the entropy explained by each division.¹⁹

Moreover, the clusters produced by any OTU picking method may also vary in homogeneity and within-cluster diversity across the various samples. Describing this within-cluster variation may lead to sub-OTU level taxonomic insights such as sharing of an OTU subpopulation across samples. One computational approach to capture this variation (termed *oligotyping*) uses Shannon entropy calculations to detect the most informative nucleotide variation and to correctly identify subpopulations within predefined OTU clusters.²⁰ This method relies on a combination of strategies to de-emphasize likely sequencing errors compared with true strain variation and has been successfully applied to track strains of *Gardnerella vaginalis* shared between sexual

partners¹¹ and to study population dynamics in the oral microbiome and in sewage.^{19,68} Another approach for extracting more detailed taxonomic information from 16S rRNA reads relies on longer read sequencing technologies (most notably, PacBio Single Molecule, Real-Time sequencing) to obtain sequence data from more variable regions of this gene. Two recently introduced pipelines process and cluster PacBio circular consensus sequencing reads,^{21,69} accounting for the specific characteristics of this different sequencing platform.

Notably, as methodologies for characterizing 16S rRNA data sets have proliferated, so have studies comparing and evaluating these approaches.^{62,70-72} These studies, however, have not necessarily reached a clear consensus on the superior approach but have rather demonstrated that the choice of algorithm can have substantial impact on subsequent analyses, and that the best choice of method likely depends on the community being analyzed and the sequencing technology used.

Once sequences have been grouped into OTUs, OTU abundances can be analyzed and compared across taxa or samples. However, accurate measurements of 16S rRNA read count may not necessarily accurately mirror the abundances of the various taxa in the community. First, because the copy number of the 16S rRNA gene varies across microbial genomes, 16S-based surveys may overestimate the abundances of taxa with multiple copies of this gene. Using reference genome information to normalize this variation can adjust and improve estimates of the relative abundance of different taxa in the same sample.^{22,73,74} Polymerase chain reaction amplification can also introduce bias into abundance comparisons between taxa, because ribosomal genes from some taxa may amplify poorly with commonly used primer sets.⁷⁵ This limitation also prevents the comparison of taxonomic abundances across different data sets generated using different primers. Finally, the *relative* abundance of reads assigned to a given OTU across samples can be skewed by changes in the *absolute* abundance of another OTU, a phenomenon known as compositional bias. A number of tools have been introduced to correct this bias, primarily by adopting techniques developed to address a similar problem in RNA-Seq experiments,²³⁻²⁵ or alternatively to account for this effect in analyzing relative abundance values.⁷⁶ Failure to address this bias can result, for example, in the identification of spurious correlations between the abundances of different OTUs, limiting our ability to robustly analyze co-occurrence relationships between different taxa.⁷⁷ Taken together, these various biases render the relationship between the relative abundances of 16S rRNA reads and true taxonomic

abundances extremely complex. One recent study set out to comprehensively characterize the joint impact of the various factors influencing this relationship by using synthetic mock communities of vaginal microbiome taxa and fitting regression models that predict true abundance of a given taxon as a function of both 16S rRNA read count and several taxon-specific bias correction terms.⁷⁸ Although such a detailed approach can be helpful for interpreting and analyzing 16S rRNA data sets of well-studied taxa, the precise relationship between 16S read counts and true community taxonomic structure for many microbiome studies remains to be characterized.

Resolving strain-level taxonomy from shotgun metagenomic data. Although 16S rRNA-based surveys can provide important insights into the taxonomic composition of a given microbiome sample, their ability to resolve strain-level genetic diversity is inherently limited. In fact, substantial genotypic variation can exist in the absence of noticeable 16S rRNA sequence divergence.⁷⁹ This variation can impact the capacity and behavior of a species and ultimately impact community-level activity. For example, some species, such as *E. coli*, have extremely marked variation in the gene content,^{79,80} which can influence the strain's pathogenicity or ecological niche.^{81,82} Moreover, the concept of a bacterial species is in fact somewhat subjective and may not be captured well by the level of divergence in a ribosomal gene sequence.⁶ It is therefore often informative to go beyond species-level resolution and to characterize the composition of strains (ie, within-species taxonomic divisions) and strain-level variation within the microbiome.

Unfortunately, however, traditional methods for detecting, characterizing, and tracking strain-level diversity rely on sequencing⁸³⁻⁸⁷ or applying microarrays^{88,89} to cultured isolates, and are therefore not readily applicable in a microbiome setting where many microbial taxa cannot be easily isolated or cultured.^{90,91} Moreover, efforts to isolate all strains of interest in a given microbiome sample and sequence their genomes can be extremely resource-intensive.⁹² An increasingly feasible alternative is to decipher strain-level diversity directly from shotgun metagenomic data using a plethora of novel and sophisticated computational techniques. Indeed, by identifying within-species genetic variation directly from metagenomic samples, a more comprehensive set of strains can be characterized in a high-throughput manner from a single sequencing experiment. This approach has been successfully applied, for example, to detect pathogenic strains of *E. coli* in clinical samples or for biosurveillance,^{26,27} to identify novel strain-level dynamics in the infant gut,²⁸ to confirm the retention of

personal strains over time,³¹ and to demonstrate extensive, widespread, and clinically relevant strain-level variation in the gut microbiome.³⁰

Notably, strain-level variation can manifest in 2 ways: single-nucleotide polymorphisms (SNPs) within shared genomic content, and variation in the presence (or copy number) of complete genes or specific segments of the genome. Most recently developed metagenomics-based SNP analysis methods take advantage of reference genome collections to estimate community diversity, detect strains of interest, or find shared strains between different metagenomic samples. These methods may use full genomes^{26,27,31,93} or marker genes known to contain loci with strain-identifying SNPs.²⁸ Because some reference genomes may be extremely similar to each other, the first step in many of these methods is to cluster genomes by similarity and select a representative genome for each cluster to which metagenomic reads can be aligned. This step speeds up the alignment process and reduces the likelihood of reads from the same strain being mapped across different but closely related references. One method, Sigma,²⁷ first uses such reference genome alignments to estimate relative abundances of taxa and then relies on the obtained estimates to refine the read-mapping assignments. ConStrains²⁸ takes a more efficient approach of mapping reads only to previously identified informative marker genes, which facilitates strain profile comparisons between samples (eg, for tracking strain dynamics in the infant microbiome^{94,95}). Another tool, PathoScope,²⁹ identifies strains using reference genomes and estimates the relative share of different strains from the same species in a given sample, and is particularly optimized to detect low-abundance strains from clinical samples.

The detection of strain-level variation in gene copy number or in gene content is a more specialized application of shotgun metagenomic-based taxonomic classification that focuses specifically on functional variation. Indeed, copy number variations are an important source of functional difference between strains, with many variable genes involved in metabolism,^{96,97} membrane and transport proteins,⁸⁷⁻⁸⁹ and virulence.^{86,96} Identifying which genes are present, absent, or vary in copy number across the various strains in a microbiome sample is therefore a crucial task that has been addressed by several recent studies. Most of these studies rely on mapping short metagenomic reads to some set of reference genomes (using a variety of read-mapping strategies), aiming to detect genomic regions for which the observed coverage varies from our expectation. Analysis of data from the Human Microbiome Project, for example, used a similar approach, mapping reads directly to a reference genome of *Strep-*

tococcus mitis and demonstrating strain-level variation in the presence/absence of various genomic elements of this species.¹ More recently, a first large-scale analysis of strain-level copy number variation was introduced, using universal single-copy genes to translate coverage measurements into copy number estimates and inferring the copy number of thousands of genes across dozens of species and in more than 100 samples.³⁰ Comparing copy number estimates across samples, this study has demonstrated extensive and widespread strain variation in the gut, including variation associated with obesity and inflammatory bowel disease. Several later studies used a similar approach for detecting strain-level variation but focused mostly on the presence/absence of genes rather than on variation in copy number.⁹⁸ Other studies extending this approach have first constructed pan-genomes (as inventories of all genes known to occur in any strain of a particular species) and mapped reads to these pan-genomes.^{31,99} An alternative approach to directly mapping short reads to reference genomes is to first assemble metagenomic reads into contigs, identify predicted genes in these contigs, and then align those to a reference.^{96,97} These longer query sequences may improve strain-specific gene identification but may be more limited in scale.

Assembling reference genomes from metagenomes. Detailed identification of strains and species can be more informative if combined with information about the gene content of each genome. Genome content information represents a mechanistic link between the taxonomy of a given microbial organism and its functional capacity, and, more generally, between community ecology and community-wide activity. Indeed, many prevalent gut species have been isolated and sequenced, yet many microbial taxa (and many strains) still lack any reference sequence.¹⁰⁰ Assembling complete genomes directly from shotgun metagenomic reads is therefore a crucial (although clearly nontrivial) task. A recent study, for example, has demonstrated the utility of assembling shotgun reads for linking taxonomic and functional dynamics after a dietary intervention for patients with Prader-Willi syndrome.¹⁰¹ The past several years have, however, witnessed substantial progress in the quality and number of genomes recovered and assembled from metagenomes.¹⁰²

Assembling genomes from metagenomes commonly involves 2 steps. First, shotgun metagenomic reads are assembled into contigs and then the obtained contigs are grouped into multiple bins such that each bin ideally includes contigs from the same taxon. The assembly step can be performed using numerous assemblers that have been optimized for assembling metagenomic reads

such as MEGAHIT,³² MetaVelvet-SL,³³ Ray Meta,¹⁰³ or IDBA-UD.¹⁰⁴ The binning step often relies on nucleotide composition, exploiting the relationship between phylogenetic relatedness and similarity in various sequence features such as GC content or *k*-mer frequency.¹⁰⁵ Such nucleotide composition-based methods are prevalent, well established, and several different implementations are available.^{106,107} More recently, a different strategy for binning was introduced, which uses the fact that different reads from the same species will tend to covary in abundance across samples.^{97,108} This approach was later refined by using the obtained differentially abundant bins of contigs to reassemble the reads.¹⁰⁹ Finally, over the past few years, several exciting methods that integrate both the nucleotide composition-based approaches and the differential abundance-based approaches have been published, including GroopM,³⁵ MaxBin,³⁷ MetaBAT,³⁶ CONCOCT,³⁴ and ABAWACA.³⁸ Another recently introduced method, termed *Latent Strain Analysis*, bins genomes using single-value decomposition, enabling it to assemble genomes from very large data sets and thus identifying rare species not found with other methods.¹¹⁰ Alternative approaches bin reads or whole genes without assembly, for example, by using the expected covariation between the abundance of various genomic elements in the metagenome and the abundance of the OTU from which they originated to deconvolve the metagenome into taxon-specific genomic data.⁴⁰ When considering these various binning methods, it should be noted that nucleotide composition-based methods have the advantage of being applicable even when only a single metagenomic sample is available, whereas differential abundance-based methods require multiple (and ideally a large number of) samples. When multiple samples are available, however, recent methods that combine both nucleotide composition and differential abundance will likely perform best. A comprehensive comparison of the performance of these many different binning algorithms has not yet been presented, although tools for validating the quality and completeness of assembled genomes are available (see, eg, CheckM³⁹ and MetaQUAST¹¹¹).

Although the methods mentioned previously relied solely on metagenomic short read data, new molecular technologies hold promise for improving metagenome-based genome assembly. For example, combining short read sequencing with Hi-C data (which provide information about the physical proximity of the different sequences) has shown to improve contig binning in synthetic mixtures of microbes.¹¹²⁻¹¹⁴ Long read and single-molecule sequencing can similarly help to link sequences from the same genome. For example, PacBio reads have been combined with short

reads to reconstruct high-quality, closed genomes from the skin microbiome,¹¹⁵ and synthetic long reads have been successfully used to improve assembly quality.^{116,117} These approaches require additional experimental and computational steps, but may significantly improve the ability to recover quality genomes from complex community samples, and are particularly promising for recovering genomes of rare species.¹¹⁷

HIGH-RESOLUTION CHARACTERIZATION OF THE MICROBIOME'S FUNCTIONAL CAPACITY

Taxonomic analyses can be extremely useful for detecting disease-associated shifts in community composition and for characterizing states of ecological dysbiosis. Some research questions, however, may be best addressed by considering the aggregate functional potential of the microbiome, regardless of the individual species that carry a specific gene or perform a specific function. Identifying which gene families are encoded in a metagenome provides insight into the capacity of the community as a whole and allows for comparison of the functional potential of a given sample to that of another sample or another environment.¹¹⁸ It can facilitate, for example, the identification of novel metabolic functions,^{119,120} disease-associated shifts in the microbiome's metabolic capacity,^{2,121} functional profile variations because of environmental fluctuations,^{122,123} or antibiotic resistance genes.^{42,124} In such settings, researchers commonly take a gene-centric approach, treating the community as a single supraorganism¹²⁵⁻¹²⁷ and profiling the set of genes collectively encoded by the metagenome. To this end, these studies directly annotate each read in the metagenome (or each gene identified in assembled contigs) with a functional category. Importantly, this approach is particularly useful when the community harbors many poorly characterized species with no reference genome. In this section, we describe recent developments in functional annotation of metagenomic samples that aim to provide a more nuanced, targeted, and accurate quantification of an individual microbiome's functional capabilities.

Accurate annotation and quantification of the metagenome's functional profile. Functional annotation of shotgun metagenomic reads can be accomplished by a variety of recently introduced frameworks¹²⁸⁻¹³³ and is typically based on mapping these reads to genes or protein domains with known functional classifications. Read mapping is done either by aligning each read to a reference database of gene or protein sequences or by using probabilistic models (such as hidden Markov models; HMMs) to evaluate

the likelihood that a given read belongs to a specific protein family or domain.

The general annotation approach provides a useful broad overview of the functional profile of a community but may have a high false positive rate because of the large reference databases used. Such false positives may represent, for example, reads originating from genes that in fact have no closely related references in the database but that still map to genes with which they share regions of homology although they may not perform the same function. To address this shortcoming, recent efforts have produced tailored reference databases that cover specific classes of proteins, in the hope that such specialized databases could improve the specificity and accuracy of functional annotation. Specifically, although large databases of protein-based HMMs exist, several specialized HMM databases have been recently introduced for metagenomic annotation. For example, FOAM⁴¹ is a database designed to identify genes matching KEGG Orthology groups^{131,134} that can aid in characterizing the metabolic potential of communities.^{135,136} Resfams,⁴² on the other hand, was developed to recognize the structure of antibiotic resistance genes and has been used to study the human gut resistomes of different cultures.^{137,138} Yet another database, dbCAN,⁴³ specifically targets carbohydrate-active enzymes. A related method, *ShortBRED*,⁴⁴ similarly quantifies a specialized set of proteins of interest, but uses alignment-based annotations rather than HMMs for a more efficient and general approach that allows for customized user-defined reference databases. These metagenomics-specific and specialized databases are a key component for accurate annotation of complex metagenomic samples. Much progress has also been made in methods for read alignment, focusing primarily on speeding up the alignment process,^{132,133} or providing efficient web-based annotation tools.^{129,131}

Notably, however, even when shotgun reads are aligned to an appropriate database, the resulting calculated functional profile can be markedly impacted by various factors, including experimental and computational biases and the protocol used to annotate each read based on the obtained alignments. Sample processing and library preparation can, for example, bias the predicted functional profile of a metagenomic sample.¹³⁹ A recent study systematically evaluated such homology-based annotation practices and demonstrated that variation introduced by computational protocol selection could completely mask true biological variation between samples, suggesting goal-specific best-practice guidelines for metagenomic annotation.¹⁴⁰ Moreover, once the samples' functional profiles have been determined, rigorous normalization and calibration of samples are still required to allow accurate comparison

across samples (eg, to identify disease-associated functional shifts). A couple of recent studies, however, have demonstrated that the commonly used compositional normalization (ie, using the *relative* abundance of each gene family within the metagenome) introduces marked biases both across and within microbiome samples.^{9,45} These studies have further presented novel methods (termed MUSiCC⁹ and MicrobeCensus⁴⁵) that use universal single-copy genes to calibrate measurements of gene abundances and to correct these biases. Use of these methods should improve the accuracy and statistical power of future comparative functional analyses.

Integrated characterization of function and taxonomy. As noted previously, methods for characterizing both the microbiome's taxonomic profile and its functional capacity have advanced rapidly over the past few years. Yet, a remaining important challenge is the integrated analysis of these 2 aspects of the microbiome and the determination of which taxa provide which functions. Such information will not only allow us to fill in gaps in the availability of reference genomes but is also a crucial first step in the development and design of targeted microbiome manipulations that could modulate the community's function.

A simple approach to associate taxa with functional potential is to annotate reads (or partial assemblies) with both taxonomy and function using any of the methods discussed previously. Determining the taxon of origin for the many reads in a metagenomic sample, however, can be both computationally expensive and methodologically challenging because of the short length of shotgun reads and varying distribution of taxonomy-distinguishing loci across genomes. To address the latter issue, early tools such as MEGAN¹⁴¹ and MTR¹⁴² used a lowest common ancestor (LCA) approach that assigns a read the highest resolution taxonomic classification that is shared by all sequences to which the read aligned. LCA classifications are clearly limited in resolution, leaving a large fraction of reads with only a course-grained taxonomic assignment or none at all.¹⁴³ To improve the precision of taxonomic assignment of shotgun metagenomic reads, several recently introduced tools have incorporated information on *k*-mer frequency profiles in reference genome databases, although how those profiles are used varies greatly between tools. LMAT⁴⁶ and Kraken⁴⁷ both assign taxonomy based on identified LCA taxa for *k*-mers in each query sequence. Other methods train models on the *k*-mer profiles associated with each taxon, using a variety of machine learning approaches including neural networks (TAC-ELM⁴⁸), naïve Bayes classifiers (RITA¹⁴⁴), or linear model-based methods.¹⁴⁵ TAC-ELM also incorporates data on GC

content and RITA combines BLAST-based reference alignments. Comparisons between Kraken and the linear model-based method mentioned previously suggest that although exact k -mer matching methods such as LMAT and Kraken are more accurate when query sequences originate from reference genomes, they may produce overly specific classifications for sequences from genomes absent from the reference database.¹⁴⁵ Moreover, Kraken requires fairly long (31 amino acid) k -mer matches, which may potentially reject many short reads because of insufficient data. These observations suggest that exact k -mer matching methods are most appropriate when a metagenome is dominated by well-characterized taxa and consists of sufficiently long reads, whereas machine learning approaches are superior for samples with more novel or unclassified microbes.

A few alternative methods for taxonomic assignment of shotgun reads use more specialized techniques. MetAnnotate⁴⁹ first uses an HMM approach to functionally annotate metagenomic reads and then determines taxonomy based on comparisons with the homologs of the matching protein family. Notably, this approach combines both functional annotation and taxonomic assignment into a single pipeline. Another tool, MetaCluster-TA,⁵¹ partially assembles reads, clusters the resulting contigs, and then assigns the LCA taxonomy given cluster alignments to genomes. SeMeta⁵⁰ similarly groups reads that contain overlapping sequence, clusters those groups by k -mer profiles, and then assigns each cluster with a taxonomic classification using an LCA approach for representative reads. These clustering-based methods aim to leverage groups of reads to obtain a broader genomic context for taxonomic classification (in contrast to the k -mer approaches that classify single reads), but may still produce low-resolution or incorrect taxonomic assignments if clustering of reads is incorrect. Together, these novel techniques allow more detailed and accurate functional profiling of microbiome samples, which will ultimately aid in understanding the human microbiome's functional capacity, dynamics, and impact on the host.

CHARACTERIZATION OF OTHER MICROBIOME FACETS VIA META-OMIC ASSAYS

Although deep genomic characterization of microbial communities has rapidly advanced our understanding of community structure and function, many community features cannot be captured by metagenomic assays. For example, the oral microbiome undergoes a dramatic shift in metabolism in response to carbohydrate consumption without any taxonomic group shifting substantially in abundance.¹⁴⁶ Likewise, communities with very different taxonomic profiles may in fact

have similar functional metabolic profiles.¹⁴⁷ To study such processes in detail and to characterize these additional facets of the microbiome's activity, researchers use comprehensive "meta-omic" technologies (including metatranscriptomics, metaproteomics, and metabolomics) that can systematically characterize community-wide gene expression, protein abundance, and metabolite concentration over time or in response to perturbations. In fact, multiple reviews have recently called for an integrative approach that combines and compares these omic assays to identify and characterize the underlying biological mechanisms in the microbiome.¹⁴⁸⁻¹⁵¹ However, analyzing each of these omic data sets presents substantial bioinformatic challenges that have only been partially addressed to date. As in metagenomics, accurate and high-resolution quantification of the measured elements and accounting for various regularities, biases, and dependencies in the data are key for realizing the full potential of these exciting high-throughput data sets. These meta-omic assays and the unique challenges each one presents are discussed subsequently.

A metatranscriptomic assay generally involves reverse transcription and complementary DNA sequencing of RNA material isolated from a microbiome sample. Such measurements of gene expression at the community level can provide important information on how different species respond to each other and to environmental changes such as antibiotic treatment¹⁵² or dietary perturbations.¹⁵³ This technology was further used to characterize gene expression patterns in a diverse range of communities.¹⁵³⁻¹⁵⁷ A typical analysis of such metatranscriptomic data consists of transcript assembly, annotation with functional and/or taxonomic information, normalization, and testing for differential expression between sample groups. None of these processing and analysis steps is necessarily simple or straightforward. The assembly of metatranscriptomic data can be performed by any transcript assembler, but it may be useful to leverage reference information from an associated metagenome. For example, a recently developed method applied a de Bruijn graph-based approach to incorporate information on metagenome assembly quality and completeness to improve subsequent transcript assembly.⁵² Assembled transcripts can be annotated for taxonomy and function using any of the metagenome annotation tools described previously. However, as in the case of metagenomic assembly, fully assembled transcripts may not always be easy to obtain or informative (although, as an alternative, an assembly free metatranscriptome-specific annotation pipeline is also available¹⁵⁸). Moreover, a recent simulation study recommended that a reasonably

unbiased analysis could be achieved by both assembling transcripts and including unassembled transcripts in subsequent clustering and annotation.¹⁵⁹ Notably, even after the metatranscriptome has been processed and the number of reads associated with each gene and/or taxon has been calculated, evaluating and exploring such data is a daunting task because of the potentially thousands of taxa, each with thousands of expressed genes, that are represented by these data. To address this challenge, an interactive tool (termed *Anvi'o*) has been recently introduced, implementing several metatranscriptomic and metagenomic processing algorithms and producing clear visualizations of assemblies and profiles at the species, gene, contig, and sample level.⁵³

Statistically sound normalization and rigorous quantitative comparisons of such complex metatranscriptomic data sets are a further challenge. The abundance of reads from a given transcript in a metatranscriptome depends on multiple factors, including the expression level of that transcript in its resident species, the abundance of that species in the community, and various biases associated with RNA-Seq experiments (such as compositional bias and batch effects). Extensive simulation and evaluation of such RNA-Seq biases and the development of rigorous methods for addressing them have produced useful tools for analyzing single-organism RNA-Seq experiments and correcting RNA-Seq-associated biases,¹⁶⁰⁻¹⁶² some of which have already been applied in the microbial community setting. In contrast, however, methods for differentiating between transcript abundance changes occurring because of gene regulation in a given taxon versus those occurring because of ecological shifts are still lacking and are an important area for future research.

Similarly, although metaproteomic assays present a powerful opportunity to understand protein-level regulation in complex communities, the analysis of such data presents a plethora of challenges, including both the traditional obstacles associated with proteomics-based experiments and additional complications associated with assaying a mixed community of microbes. Such studies generally use tandem mass spectrometry to quantify peptide fragments and then identify the source proteins of each peptide by searching against a reference database of theoretical or previously collected spectra. Because a peptide typically cannot be identified unless it is found in the reference database, the choice of database and search parameters can have a substantial impact on the obtained results. Indeed, this effect was convincingly shown in a recent study comparing peptide identifications in a human intestinal metaproteomic data set with a classic single-organism proteomic data set using several different metaproteomic databases and search strategies.¹⁶³ An efficient way to narrow the

search space and identify uncharacterized proteins is to use a database of theoretical spectra constructed from associated metagenome sequencing reads to search the obtained peptides.¹⁶⁴ In a recent study, for example, a strain-resolved metagenome was used to analyze a longitudinal metaproteomic data set from the gut of a preterm infant.¹⁶⁵ Furthermore, difficulty associated with analyzing community-wide proteomic data arises because a given peptide may match homologous proteins across multiple taxa. *Pipasic* is a recently developed tool that addresses this challenge by correcting for the amount of similarity in peptide sequences from different strains.⁵⁴ Moreover, proteins at the community level display an enormous dynamic range of abundances, and it therefore cannot be reliably determined whether a peptide not detected in a given sample by an untargeted assay is indeed completely absent or present but at a very low abundance. This incompleteness restricts the utility of metaproteomics for community metabolism modeling, although this limitation may be ultimately mitigated by improving technology. As a promising example, one recent study was able to use metaproteomic data to construct and compare detailed metabolic models of 2 naphthalene-degrading bacterial communities.¹⁶⁶

Importantly, although genes and proteins vary across taxa, metabolites are, at least in principle, universal. Accordingly, in contrast to metatranscriptomics and metaproteomics, the processing and analysis of community-wide metabolomic data can rely on standard approaches for single-organism metabolomics with essentially no modifications. For untargeted mass spectrometry metabolomics, these analyses typically involve normalization and putative identification of metabolites by searching either for matches in a spectral library or for known compounds with matching mass and chromatographic elution profiles.¹⁶⁷ The greater challenge, however, lies in the interpretation of these data sets and in linking the observed variation in biomolecule abundances with other data on community structure and function. Statistical associations between disease, metabolite concentrations, and microbial species abundances have been observed in case-control studies of Crohn's disease, colorectal cancer, and *Clostridium difficile* infection among other conditions, but the mechanistic nature of these links remains unclear.¹⁶⁸⁻¹⁷¹ A few studies have further used metabolic pathway information to quantify the link between shifts in the metagenome and functionally related metabolome variation.^{172,173} Moreover, a recent study has introduced a novel computational framework, MIMOSA, for metabolic model-based integration of community taxonomic and metabolomic data and for evaluating whether variation in the metabolome can

be explained mechanistically by variation in the community's taxonomic profile.⁵⁶ Such methods are crucial for gaining a principled, system-level understanding of how changes in community ecology impact community metabolism and behavior.

Finally, although not strictly an omic assay, high-resolution imaging of microbial communities and the study of community spatial distributions are another area of rapid technology and bioinformatic development. Spatial factors can affect microbial community nutrient availability, communication, and biofilm formation, among other processes.¹⁷⁴ Methods for quantifying the distribution of microbes in a community and relating it to associated omic data are therefore clearly needed. One increasingly popular technique is fluorescent *in situ* hybridization with primers specific to various bacterial taxa of interest, combined with high-resolution microscopy.^{55,175} A recently developed tool, called BacSpace, systematically processes and analyzes such data by filtering out nonmicrobial fluorescence and calculating and aggregating distances between different microbial cells and environmental landmarks.⁵⁵ Another approach to examine microbial biogeography on a larger scale involves mapping and visualizing many metabolomic and taxonomic profiles via a 3D model of a community site. This strategy has been applied to communities growing on solid culture¹⁷⁶ and to the human skin microbiome.¹⁷⁷ Computational and quantitative methods along these lines are crucial for incorporating information of spatial heterogeneity into a more complete mechanistic and quantitative understanding of the microbiome.

CONCLUSIONS

The growing appreciation for the scientific and clinical importance of the human microbiome has given rise to an explosion of microbiome studies. These studies now routinely generate, assemble, and explore high-dimensional meta-omic data at an unprecedented scale. Previously, we have broadly outlined the most common types of approaches and computational tools available for processing and analyzing such data, with emphasis on several areas in which increasingly higher resolution and precision can be gained from computational analysis of microbiome data (Fig 1). Fortunately, such tools are regularly distributed as open-source software that can be applied to data sets from a wide range of studies (Table 1). It is important to note, however, that these methods (and likely many other methods that will be developed to address these challenges in coming years) are ultimately limited by the large number of genes of unknown function and yet-uncharacterized taxa present in the microbiome. Developing efficient,

cost-effective, and rigorous methods to demystify these hidden layers of microbiome diversity is therefore necessary to realize the full potential of microbiome research. Nevertheless, the resolution and scale of microbial community profiling will likely continue to improve with future technology development. These technologies will provide an increasingly more detailed view of the structure and function of the microbiome's subpopulations and even single cells across time and space, the behavior of such subpopulations, and the way they interact with one another and with the host. These advances will contribute to the growing field of personalized medicine, with applications ranging from precise identification of pathogenic strains for targeted treatment, through careful monitoring of dysbiotic microbial communities in disease, to personalized and rational design of microbiome manipulations.

ACKNOWLEDGMENTS

Conflicts of Interest: All authors have read the journal's policy on disclosure of potential conflicts of interest. All authors have disclosed any financial or personal relationship with organizations that could potentially be perceived as influencing the described research.

C.N. is supported by an NSF IGERT DGE-1258485 fellowship. C.P.M. is supported by "Interdisciplinary Training in Genomic Sciences" National Human Genome Research Institute Grant T32 HG00035. This work was supported in part by New Innovator Award DP2 AT007802-01 to E.B.

All authors have read the journal's authorship agreement and the manuscript has been reviewed and approved by all authors.

REFERENCES

1. Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; 486:207–14.
2. Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60.
3. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JL. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 2006;444:1027–31.
4. Cox LM, Yamanishi S, Sohn J, et al. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell* 2014;158:705–21.
5. Smith MI, Yatsunenko T, Manary MJ, et al. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 2013; 339:548–54.
6. Yarza P, Yilmaz P, Pruesse E, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 2014;12:635–45.
7. Shakya M, Quince C, Campbell JH, Yang ZK, Schadt CW, Podar M. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ Microbiol* 2013;15:1882–99.

8. Mande SS, Mohammed MH, Ghosh TS. Classification of metagenomic sequences: methods and challenges. *Brief Bioinform* 2012;13:669–81.
9. Manor O, Borenstein E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* 2015;16:53.
10. Manor O, Levy R, Borenstein E. Mapping the inner workings of the microbiome: genomic- and metagenomic-based study of metabolism and of metabolic interactions in the human gut microbiome. *Cell Metab* 2014;20:742–5.
11. Eren AM, Zozaya M, Taylor CM, Dowd SE, Martin DH, Ferris MJ. Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *Ravel J*, ed. *PLoS One* 2011;6:e26732.
12. Fitz-Gibbon S, Tomida S, Chiu BH, et al. Propionibacterium acnes strain populations in the human skin microbiome associated with acne. *J Invest Dermatol* 2013;133:2152–60.
13. Busby B, Kristensen DM, Koonin EV. Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environ Microbiol* 2012;15:307–12.
14. Haiser HJ, Gootenberg DB, Chatman K, Sirasani G, Balskus EP, Turnbaugh PJ. Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* 2013;341:295–8.
15. Hajishengallis G, Liang S, Payne M, et al. Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement. *Cell Host Microbe* 2011;10:497–506.
16. Charbonneau M, O'Donnell D, Blanton L, et al. Sialylated milk oligosaccharides promote microbiota-dependent growth in models of infant undernutrition. *Cell* 2016;164:859–71.
17. Wang X, Yao J, Sun Y, Mai V. M-pick a modularity-based method for OTU picking of 16S rRNA sequences. *BMC Bioinformatics* 2013;14:43.
18. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 2015;3:e1420.
19. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* 2014;9:968–79.
20. Eren AM, Maignien L, Sul WJ, et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* 2013;4:1111–9.
21. Franzén O, Hu J, Bao X, Itzkowitz SH, Peter I, Bashir A. Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. *Microbiome* 2015;3:43.
22. Angly FE, Dennis PG, Skarshewski A, Vanwonderghem I, Hugenholtz P, Tyson GW. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* 2014;2:11.
23. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;10:1200–2.
24. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* 2014;10:e1003531.
25. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* 2015;31:2269–75.
26. Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P. Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. *Genome Med* 2015;7:52.
27. Ahn TH, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 2014;31:170–7.
28. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* 2015;33:1045–52.
29. Hong C, Manimaran S, Shen Y, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* 2014;2:33.
30. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* 2015;160:583–94.
31. Nayfach S, Pollard KS. Population genetic analyses of metagenomes reveal extensive strain-level variation in prevalent human-associated bacteria. *bioRxiv* 2015;031757.
32. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;31:1674–6.
33. Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res* 2014;22:69–77.
34. Alneberg J, Bjarnason BS, de Bruijn I, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;11:1144–6.
35. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2014;2:e603.
36. Kang DD, Froula J, Egan R, Wang Z. MetaBAT an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015;3:e1165.
37. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2015;32:605–7.
38. Brown CT, Hug LA, Thomas BC, et al. Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* 2015;523:208–11.
39. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates single cells, and metagenomes. *Genome Res* 2015;25:1043–55.
40. Carr R, Shen-Orr SS, Borenstein E. Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. *PLoS Comput Biol* 2013;9:e1003292.
41. Prestat E, David MM, Hultman J, et al. FOAM (functional ontology assignments for metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res* 2014;42:e145.
42. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 2015;9:207–16.
43. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* 2012;40:W445–51.
44. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. *Noble WS*, ed. *PLoS Comput Biol* 2015;11:e1004557.
45. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* 2015;16:51.

46. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* 2013;29:2253–60.
47. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014;15:R46.
48. Rasheed Z, Rangwala H. Metagenomic taxonomic classification using extreme learning machines. *J Bioinform Comput Biol* 2012;10:1–19.
49. Petrenko P, Lobb B, Kurtz DA, Neufeld JD, Doxey AC. MetAnnotate: function-specific taxonomic profiling and comparison of metagenomes. *BMC Biol* 2015;13:92.
50. Le VV, Tran LV, Tran HV. A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads. *BMC Bioinformatics* 2016;17:22.
51. Wang Y, Leung H, Yiu S, Chin F. MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics* 2014;15:S12.
52. Ye Y, Tang H. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics* 2016;32:1001–8.
53. Eren AM, Esen ÖC, Quince C, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015;3:e1319.
54. Penzlin A, Lindner MS, Doellinger J, Dabrowski PW, Nitsche A, Renard BY. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics* 2014;30:i149–56.
55. Earle K, Billings G, Sigal M, et al. Quantitative imaging of gut microbiota spatial organization. *Cell Host Microbe* 2015;18:478–88.
56. Noecker C, Eng A, Srinivasan S, et al. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 2016;1:e00013–5.
57. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a microbiome study. *Cell* 2014;158:250–62.
58. Kostic AD, Gevers D, Pedamallu CS, et al. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res* 2012;22:292–8.
59. Llopis M, Cassard AM, Wrzosek L, et al. Intestinal microbiota contributes to individual susceptibility to alcoholic liver disease. *Gut* 2016;65:830–9.
60. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci* 1977;74:5088–90.
61. Ju F, Zhang T. 16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions. *Appl Microbiol Biotechnol* 2015;99:4119–29.
62. Kopylova E, Navas-Molina JA, Mercier C, et al. Open-source sequence clustering methods improve the state of the art. *mSystems* 2016;1:e00003–15.
63. Tikhonov M, Leach RW, Wingreen NS. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *ISME J* 2014;9:68–80.
64. Forster D, Bittner L, Karkar S, et al. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol* 2015;13:16.
65. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2014;2:e593.
66. De Vargas C, Audic S, Henry N, et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015;348:1261605.
67. Lima-Mendez G, Faust K, Henry N, et al. Determinants of community structure in the global plankton interactome. *Science* 2015;348:1262073.
68. Newton RJ, McLellan SL, Dila DK, et al. Sewage reflects the microbiomes of human populations. *MBio* 2015;6:e02574–614.
69. Singer E, Bushnell B, Coleman-Derr D, et al. High-resolution phylogenetic microbial community profiling. *ISME J* 2016;10:2020–32.
70. Westcott SL, Schloss PD. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* 2015;3:e1487.
71. Forster D, Dunthorn M, Stoeck T, Mahé F. Comparison of three clustering approaches for detecting novel environmental microbial diversity. *PeerJ* 2016;4:e1692.
72. Schmidt TSB, Rodrigues JFM, von Mering C. Ecological consistency of SSU rRNA-based operational taxonomic units at a global scale. *PLoS Comput Biol* 2014;10:e1003594.
73. Kembel SW, Wu M, Eisen JA, Green JL. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput Biol* 2012;8:e1002743.
74. Langille MGI, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013;31:814–21.
75. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* 2015;3:26.
76. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. Mering C, ed. *PLoS Comput Biol* 2012;8:e1002687.
77. Weiss S, Van Treuren W, Lozupone C, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J* 2016;10:1669–81.
78. Brooks JP, Edwards DJ, Harwich MD, et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* 2015;15:66.
79. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. *Philos Trans R Soc B Biol Sci* 2006;361:1929–40.
80. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb Ecol* 2010;60:708–20.
81. Clermont O, Bonacorsi S, Bingen E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* 2000;66:4555–8.
82. LeBlanc J. Implication of virulence factors in *Escherichia coli* O157:H7 pathogenesis. *Crit Rev Microbiol* 2003;29:277–96.
83. Holt KE, Parkhill J, Mazzoni CJ, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* 2008;40:987–93.
84. Gutacker MM, Smoot JC, Migliaccio CAL, et al. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* 2002;162:1533–43.
85. Manning SD, Motiwala AS, Springman AC, et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci* 2008;105:4868–73.
86. Gill SR, Fouts DE, Archer GL, et al. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J Bacteriol* 2005;187:2426–38.

87. Hansen EE, Lozupone CA, Rey FE, et al. Pan-genome of the dominant human gut-associated archaeon *Methanobrevibacter smithii*, studied in twins. *Proc Natl Acad Sci* 2011;108:4599–606.
88. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci* 2000;97:14668–73.
89. Siezen RJ, Tzeneva VA, Castioni A, et al. Phenotypic and genomic diversity of *Lactobacillus plantarum* strains isolated from various environmental niches. *Environ Microbiol* 2010;12:758–73.
90. Rappé MS, Giovannoni SJ. The uncultured microbial majority. *Annu Rev Microbiol* 2003;57:369–94.
91. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* 2002;3:1–8.
92. Atarashi K, Tanoue T, Oshima K, et al. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* 2013;500:232–6.
93. Schloissnig S, Arumugam M, Sunagawa S, et al. Genomic variation landscape of the human gut microbiome. *Nature* 2013;493:45–50.
94. Vatanen T, Kostic AD, d’Hennezel E, et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* 2016;165:842–53.
95. Yassour M, Vatanen T, Siljander H, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med* 2016;8:343ra81.
96. Morowitz MJ, Deneff VJ, Costello EK, et al. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. *Proc Natl Acad Sci* 2010;108:1128–33.
97. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 2012;23:111–20.
98. Zhu A, Sunagawa S, Mende DR, Bork P. Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol* 2015;16:82.
99. Scholz M, Ward DV, Pasolli E, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 2016;13:435–8.
100. Hug LA, Baker BJ, Anantharaman K, et al. A new view of the tree of life. *Nat Microbiol* 2016;1:16048.
101. Zhang C, Yin A, Li H, et al. Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine* 2015;2:968–84.
102. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 2016;4:8.
103. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 2012;13:R122.
104. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 2011;27:i94–101.
105. Pride DT. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 2003;13:145–58.
106. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* 2012;3:410.
107. Saeed I, Tang SL, Halgamuge SK. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res* 2011;40:e34.
108. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013;31:533–8.
109. Nielsen HB, Almeida M, Juncker AS, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* 2014;32:822–8.
110. Cleary B, Brito IL, Huang K, et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol* 2015;33:1053–60.
111. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;32:1088–90.
112. Burton J, Liachko I, Dunham M, Shendure J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* 2014;4:1339–46.
113. Marbouty M, Cournac A, Flot JF, Marie-Nelly H, Mozziconacci J, Koszul R. Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife* 2014;3:e03318.
114. Beitel CW, Froenicke L, Lang JM, et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* 2014;2:e415.
115. Tsai YC, Conlan S, Deming C, et al. Resolving the complexity of human skin metagenomes using single-molecule sequencing. *MBio* 2016;7:e01948–2015.
116. Sharon I, Kertesz M, Hug LA, et al. Accurate multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res* 2015;25:534–43.
117. Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* 2015;34:64–9.
118. Tringe SG, von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. *Science* 2005;308:554–7.
119. Illegheems K, Weckx S, De Vuyst L. Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. *Food Microbiol* 2015;50:54–63.
120. White RA, Chan AM, Gavelis GS, et al. Metagenomic analysis suggests modern freshwater microbialites harbor a distinct core microbial community. *Front Microbiol* 2016;6:1531.
121. Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* 2012;109:594–9.
122. Freedman ZB, Upchurch RA, Zak DR, Cline LC. Anthropogenic N deposition slows decay by favoring bacterial metabolism: insights from metagenomic analyses. *Front Microbiol* 2016;7:259.
123. Koenig JE, Spor A, Scalfone N, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* 2011;108(Suppl 1):4578–85.
124. Li B, Yang Y, Ma L, et al. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* 2015;9:2490–502.
125. Lederberg J. Infectious history. *Science* 2000;288:287–93.
126. Gordon JI, Klaenhammer TR. A rendezvous with our microbes. *Proc Natl Acad Sci U S A* 2011;108:4513–5.
127. Borenstein E. Computational systems biology and in silico modeling of the human microbiome. *Brief Bioinform* 2012;13:769–80.

128. Meyer F, Paarmann D, D'Souza M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;9:386.
129. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 2011;12:444.
130. Arumugam M, Harrington ED, Foerster KU, Raes J, Bork P. SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* 2010;26:2977–8.
131. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* 2015;428:726–31.
132. Bose T, Haque MM, Reddy C, Mande SS. COGNIZER: a framework for functional annotation of metagenomic datasets. *PLoS One* 2015;10:e0142102.
133. Kultima JR, Coelho LP, Forslund K, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 2016 [Epub ahead of print].
134. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;28:27–30.
135. Yeoh YK, Paungfoo-Lonhienne C, Dennis PG, et al. The core root microbiome of sugarcane cultivated under varying nitrogen fertilizer application. *Environ Microbiol* 2016;18:1338–51.
136. Nelson MB, Berlemont R, Martiny AC, Martiny JBH. Nitrogen cycling potential of a grassland litter microbial community. *Kostka JE, ed. Appl Environ Microbiol* 2015;81:7012–22.
137. Clemente JC, Pehrsson EC, Blaser MJ, et al. The microbiome of uncontacted Amerindians. *Sci Adv* 2015;1(3):e1500183.
138. Rampelli S, Schnorr SL, Consolandi C, et al. Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr Biol* 2015;25:1682–93.
139. Jones MB, Highlander SK, Anderson EL, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci* 2015;112:14024–9.
140. Carr R, Borenstein E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PLoS One* 2014;9:e105776.
141. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res* 2007;17:377–86.
142. Gori F, Folino G, Jetten MSM, Marchiori E. MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. *Bioinformatics* 2010;27:196–203.
143. Kunitz V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A Bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* 2008;72:557–78.
144. MacDonald NJ, Parks DH, Beiko RG. Rapid identification of high-confidence taxonomic assignments for metagenomic data. *Nucleic Acids Res* 2012;40:e111.
145. Vervier K, Mahé P, Tournoud M, Veyrieras JB, Vert JP. Large-scale machine learning for metagenomics sequence classification. *Bioinformatics* 2016;32:1023–32.
146. Edlund A, Yang Y, Yooseph S, et al. Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. *ISME J* 2015;9:2605–19.
147. Ferrer M, Ruiz A, Lanza F, et al. Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ Microbiol* 2013;15:211–26.
148. Franzosa EA, Hsu T, Sirota-Madi A, et al. Sequencing and beyond: integrating molecular “omics” for microbial community profiling. *Nat Rev Microbiol* 2015;13:360–72.
149. Lamendella R, VerBerkmoes N, Jansson JK. ‘Omics’ of the mammalian gut – new insights into function. *Curr Opin Biotechnol* 2012;23:491–500.
150. Waldor MK, Tyson G, Borenstein E, et al. Where next for microbiome research? *PLoS Biol* 2015;13:e1002050.
151. Greenblum S, Chiu H, Levy R, Carr R, Borenstein E. Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Curr Opin Biotechnol* 2013;24:810–20.
152. Pérez-Cobas AE, Gosalbes MJ, Friedrichs A, et al. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut* 2013;62:1591–601.
153. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 2014;505:559–63.
154. De Filippis F, Genovese A, Ferranti P, Gilbert JA, Ercolini D. Metatranscriptomics reveals temperature-driven functional changes in microbiome impacting cheese maturation rate. *Sci Rep* 2016;6:21871.
155. Shi W, Moon CD, Leahy SC, et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. *Genome Res* 2014;24:1517–25.
156. Jorh P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the human oral microbiome during health and disease. *MBio* 2014;5:e01012–4.
157. Aylward FO, Eppley JM, Smith JM, Chavez FP, Scholin CA, DeLong EF. Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc Natl Acad Sci* 2015;112:5443–8.
158. Leimena MM, Ramiro-Garcia J, Davids M, et al. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics* 2013;14:530.
159. Toseland A, Moxon S, Mock T, Moulton V. Metatranscriptomes from diverse microbial communities: assessment of data reduction techniques for rigorous annotation. *BMC Genomics* 2014;15:901.
160. Dillies MA, Rau A, Aubert J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2012;14:671–83.
161. Qin LX, Huang HC, Niu Y. Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software. *Cancer Inform* 2015;14:57–67.
162. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
163. Muth T, Kolmeder CA, Salojärvi J, et al. Navigating through metaproteomics data: A logbook of database searching. *Proteomics* 2015;15:3439–53.
164. Erickson AR, Cantarel BL, Lamendella R, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 2012;7:e49138.
165. Brooks B, Mueller RS, Young JC, Morowitz MJ, Hettich RL, Banfield JF. Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. *Front Microbiol* 2015;6:654.
166. Tobalina L, Bargiela R, Pey J, et al. Context-specific metabolic network reconstruction of a naphthalene-degrading bacterial community guided by metaproteomic data. *Bioinformatics* 2015;31:1771–9.
167. Ren S, Hinzman AA, Kang EL, Szczesniak RD, Lu LJ. Computational and statistical analysis of metabolomics data. *Metabolomics* 2015;11:1492–513.

168. Jansson J, Willing B, Lucio M, et al. Metabolomics reveals metabolic biomarkers of Crohn's disease. *PLoS One* 2009;4:e6386.
169. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS One* 2013;8:e70803.
170. Theriot CM, Koenigsnecht MJ, Carlson PE, et al. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to *Clostridium difficile* infection. *Nat Commun* 2014;5:3114.
171. Gomez A, Petzelkova K, Yeoman CJ, et al. Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (*Gorilla gorilla gorilla*) reflect host ecology. *Mol Ecol* 2015;24:2551–65.
172. McHardy IH, Goudarzi M, Tong M, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* 2013;1:17.
173. Sridharan GV, Choi K, Klemashevich C, et al. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat Commun* 2014;5:5492.
174. Donaldson GP, Lee SM, Mazmanian SK. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* 2016;14:20–32.
175. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. *Proc Natl Acad Sci* 2016;113:E791–800.
176. Watrous JD, Phelan VV, Hsu CC, et al. Microbial metabolic exchange in 3D. *ISME J* 2013;7:770–80.
177. Bouslimani A, Porto C, Rath CM, et al. Molecular cartography of the human skin surface in 3D. *Proc Natl Acad Sci* 2015; 112:E2120–9.