HHS Public Access

Author manuscript

Nat Rev Microbiol. Author manuscript; available in PMC 2016 March 21.

Published in final edited form as:

Nat Rev Microbiol. 2015 June ; 13(6): 360–372. doi:10.1038/nrmicro3451.

Sequencing and beyond: integrating molecular 'omics for microbial community profiling

Eric A. Franzosa[1,2], **Tiffany Hsu**[1,2], **Alexandra Sirota-Madi**[2,3], **Afrah Shafquat**[1], **Galeb Abu-Ali**[1], **Xochitl C. Morgan**[1,2], and **Curtis Huttenhower**[1,2]

Curtis Huttenhower: chuttenh@hsph.harvard.edu

[1]Biostatistics Department, Harvard School of Public Health, Boston, MA 02115, USA

[2]The Broad Institute, Cambridge, MA 02142, USA

[3]Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

Abstract

High-throughput DNA sequencing has proven invaluable for investigating diverse environmental and host-associated microbial communities. In this Review, we discuss emerging strategies for microbial community analysis that complement and expand traditional metagenomic profiling. These include novel DNA sequencing strategies for identifying strain-level microbial variation and community temporal dynamics; measuring additional multi'omic data types that better capture community functional activity, such as transcriptomics, proteomics, and metabolomics; and combining multiple forms of multi'omic data in an integrated framework. We highlight studies in which the multi'omics approach has led to improved mechanistic models of microbial community structure and function.

Research in microbial community ecology has expanded enormously in the era of high-throughput functional genomics. This trend is due in large part to advances in DNA sequencing, which now enable researchers to probe microbial community composition and function in a high-resolution and culture-independent manner. In a technique called metagenomics[1], shotgun sequencing methods are applied to millions of random genomic fragments sampled from a microbial community. The resulting DNA sequence data are then typically used to assess the community in at least two ways: taxonomic profiling, which answers, "who is present in the community?" and functional profiling, which answers, "what could they be doing?" (Box 1). Another common culture-independent method for profiling a microbial community involves sequencing specific microbial amplicons (predominantly the bacterial 16S rRNA gene). Although amplicon-based sequencing considers only one or a few microbial genes, it is frequently grouped under the umbrella of metagenomics as one way to perform taxonomic, phylogenetic or functional profiling (Box 1).

Whole metagenome shotgun (WMS) and amplicon sequencing have been applied to study diverse microbiomes, ranging from natural environments[2-4] to the built environment[5] and

Correspondence to: Curtis Huttenhower, chuttenh@hsph.harvard.edu.

the human body[6,7]. For example, metagenomic profiling is applied to study shifts in human microbiome composition and function associated with human diseases, including obesity[8-10], inflammatory bowel disease[11-13], and cancer[14-16]. While these approaches to profiling microbial community structure and function have proven highly informative, current DNA sequence-based methods have limitations. For example, the most common approaches provide at best species-level taxonomic resolution, whereas many important phenomena occur at the strain level (e.g. acquisition of antibiotic resistance genes). Similarly, most common models for microbiome study design involve cross-sectional or case-control sampling, but not longitudinal sampling, and hence fail to capture the dynamic behavior of microbial communities. Addressing these issues requires new considerations at the experimental design phase, such as assessing the trade-offs between the number of environments considered (sample size, N), the depth of sequencing per environment (as greater depth facilitates strain-level analysis), and the number of time points considered per environment (Figure 1). In addition, leveraging the respective strengths of amplicon sequencing – which has lower resolution but is cheaper, – and WMS sequencing – which provides higher resolution but at a higher cost –through tiered study designs can further push the limits of what is possible with metagenomic sequencing (Figure 1).

Metagenomic sequencing faces a fundamental limitation in its inability to directly measure the functional activity of a community under a given set of conditions. Thus, additional multi'omic data are required to fully describe a microbial community, such as community RNA (transcriptomics), protein (proteomics), and metabolite abundances (metabolomics), preferably in an integrated framework. In this Review, we discuss these new directions in microbiome research and highlight examples of next-generation metagenomics and integrated multi'omics that have led to more advanced hypotheses, mechanisms, and models of microbial community evolution and function.

New approaches in taxonomic profiling

The most common limitations of traditional metagenomic analysis are the limited taxonomic resolution, which is usually restricted to the species-level, and the lack of temporal resolution. However, new strategies are emerging that allow the study of strain-level variation and the dynamic behavior of microbial communities.

Profiling strain-level variation

Typical approaches for taxonomic profiling of microbial communities do not capture strain-level variation, yet this information is crucial for accurately characterizing individual microorganisms (and by extension, communities). For example, *Escherichia coli* commonly occurs as a commensal organism in the human gut[17]; however, the acquisition of genes encoding Shiga toxin result in a subset of *E. coli* strains becoming highly pathogenic, such as the well-known serotype O157:H7[18]. Therefore, strain-level profiling, in particular profiling of gene content based on WMS or single-cell sequencing, is needed to identify such variation in uncultured or unknown organisms (Figure 2).

Taxonomic profiles based on standard amplicon sequencing are composed of Operational Taxonomic Units (OTUs), which are often more specific than genera but in the substantial

majority of cases less specific than species (Box 1). Recently, a new strategy has been proposed that uses a sequence entropy-based approach to identify maximally informative sites within the 16S rRNA gene to improve OTU resolution[19]. This strategy, called oligotyping, is advantageous for distinguishing closely related taxa (such as those that differ by a single 16S rRNA nucleotide) and has been applied to study subspecies-level population structure in the vaginal microbiome[20] and to link sewage samples to specific fecal pollution sources[21]. In addition, a new, low-error approach to 16S rRNA gene sequencing, termed LEA-Seq has been proposed and used to profile stable carriage of host-specific strains in the human gut microbiome[22].

Despite recent advances in amplicon sequencing, WMS sequencing is the preferred method for strain-level profiling due to its ability to identify variation throughout microbial genomes (Figure 2). Mapping sequences obtained by WMS sequencing (termed reads to bacterial reference genomes or sets of species-specific marker genes provides a straightforward method for profiling species composition. However, due to strain-specific gene loss events, portions of these reference sequences may be absent in isolates of a species present in the sample, resulting in gaps in otherwise uniform coverage of the reference (Figure 2). For example, mapping WMS reads from tongue samples to genomes of *Streptococcus mitis* highlighted the presence and absence of genomic islands in isolates of that species from individuals enrolled in the Human Microbiome Project (HMP)[7]. Genomic islands were shown to contain multiple, functionally coherent genes (such as subunits of the V-type H^+ ATPase) that were gained and lost together, suggesting a mechanism for individual- and body site-specific functional specialization. Profiling this type of strain variation via the presence and absence of species-specific marker genes[23] has been similarly applied to identify strains of *Prevotella copri* associated with susceptibility to arthritis[24] and to characterize the transit of abundant human oral strains to the gut[25] (Figure 2).

While missing genomic elements are detectable at relatively low WMS sequencing depths, greater depths enable confident detection of a wider variety of strain-level variants, including single nucleotide polymorphisms (SNPs; Figure 2). For example, existing WMS data from human stool samples have been used to identify reference genomes with high sequencing coverage that were then scanned for SNPs[26]. This analysis revealed that subject-specific SNP variation tended to remain stable for up to a year and was comparatively more conserved than overall species abundance. It was also possible to rank species and genes in the gut by the degree of polymorphism across individuals, which revealed that antimicrobial resistance genes were among the most variable, while housekeeping genes were among the most conserved.

In addition to allowing SNP identification, deeper WMS sequencing can also facilitate the *de novo* assembly of contigs and whole microbial genomes from metagenomes; these assembly-based approaches are particularly relevant for studying microbial communities that are poorly represented in catalogs of microbial reference genomes (Box 1). Indeed, it is increasingly possible to assemble whole microbial genomes from such communities and analyze their strain-level variation[27-31], a process that was until recently only feasible in low-complexity communities[2]. Complementing gene profiles and SNPs, assemblies can

reveal novel genomic rearrangements and horizontal gene transfer (HGT) events more readily than reference genome-based approaches (Box 1, Figure 2).

Time course analysis

The composition of microbial communities can change dramatically over time, highlighting the need for temporal profiling in order to incorporate the (sometimes substantial) longitudinal dynamics of microbial communities into analyses. For example, high temporal resolution 16S rRNA gene sequencing has been used to assess the stability of the human gut, oral, and skin microbiomes[32]. Over a time scale of approximately one year, these communities tended to maintain small, stable core members and non-core members that persisted for variable periods. Tracking microbiome development in human infants is another topic of great interest, particularly in cases where normal development is disrupted by medical intervention in early life[33]. For example, longitudinal WMS sequencing of an infant delivered by C-section revealed an early gut microbiome dominated by skin-associated microorganisms; however, the metabolic environment of the infant gut appeared to select against these early colonizers during the first months of life[34].

Longitudinal analysis is also advantageous for studying microbial community perturbations in human diseases. Indeed, such perturbations may signal the onset or progression of a disease and could serve as important biomarkers. For example, longitudinal 16S rRNA gene analysis of the human skin microbiome has been performed in children with atopic dermatitis[35], revealing increases in particular taxa associated with disease flares, including *Staphylococcus aureus* (a known correlate of atopic dermatitis) and *Staphylococcus epidermidis* (a skin commensal); changes in *S. aureus* abundance correlated with disease severity. Longitudinal sampling also highlighted the effects of treatment for atopic dermatitis, which showed that an increase in bacterial diversity occurs before the resolution of symptoms. Longitudinal approaches have been similarly applied to study the resolution of *Clostridium difficile* infection following faecal transplantion[36] (by amplicon sequencing) and to link changes in host diet to altered gut microbial composition[37,38] (by both amplicon and WMS sequencing methods). Notably, the latter examples support a role for the microbiome in shaping, and perhaps as treatment for, metabolic disorders.

Longitudinal studies are equally relevant for studying the dynamics of microbiomes outside of the human body (Figure 2). For example, one study explored the interplay between viral and microbial populations in human-controlled aquatic environments (aquaculture and solar saltern ponds)[39]. Theoretical models predict that such communities should follow “Kill-the-Winner” dynamics; as a microbial species becomes more dominant, its interactions with predatory phages increase, ultimately leading to population decline. The cycle then repeats for the next microbial species rising to dominance, always driving the community away from a homogeneous state. Contrary to this model, earlier empirical observations had shown that similar communities maintained surprisingly stable composition and metabolic potential. By using temporal metagenomic analysis, this apparent paradox was resolved by demonstrating that although composition remained stable at the species level, distinct microbial strains within those species displayed “Kill-the-Winner” dynamics, as predicted by the theoretical model. Therefore, while the net abundance of strains within a species

remained stable, individual strains grew or declined according to strain-specific phage predation. These findings highlight the advantages of integrating strain-level profiling with longitudinal sampling and serve as a reminder of the benefits of considering alternative metagenomic sequencing strategies (Figure 1).

Multi'omics analyses

The preceding sections demonstrated that DNA sequence information can be used to profile microbial communities in several insightful but underutilized ways. However, although the genomic content of a community describes its *functional potential* (what the community is capable of doing), it does not provide any information on its *functional activity* (what the community is doing in a particular condition or timepoint). The extent to which functional potential dictates functional activity in microbial communities is not well understood; by one estimate, half of the variation in functional activity in the human gut microbiome under baseline conditions is explained by functional potential (gene copy number)[25], suggesting that the remaining variation must be due to other factors (such as gene regulation). To fully understand the determinants of function, additional multi'omics data types such as transcriptomics, proteomics and metabolomics are needed.

Measuring functional activity with metatranscriptomics

Metatranscriptomics involves the sequencing of total RNA within a microbial community. It is critical to enrich for microbial mRNAs by depleting rRNA prior to metatranscriptomic sequencing, as mRNAs are dwarfed in abundance by bacterial rRNAs in the total microbial RNA pool[40]. Microbial mRNA is then converted to cDNA and sequenced by standard methods. With appropriate barcoding of DNA and cDNA samples, metagenomic and metatranscriptomic (meta'omic) sequencing can be carried out in tandem, making RNA sequencing a natural extension for microbial community surveys[40], and a further consideration during the design of sequencing-based surveys of microbial communities (Figure 1).

Metatranscriptomic approaches were first applied to freshwater and marine microbial communities[41-43]. These studies demonstrated that, like DNA, microbial total RNA could be used to profile community structure, function and diversity. Moreover, these studies showed that RNA sequencing produced large amounts of novel sequence information, presumably by capturing organisms or genes of low copy number that are under-sampled by DNA sequencing alone. In addition, metatranscriptomic sequencing also provides a means to detect and quantify RNA viruses[44,45] which are otherwise not included in DNA-based metagenomic surveys (aside from integrated retroviruses[46]).Combining metatranscriptomics with DNA-based taxonomic and functional profiling reveals prominent over- or under-expression of particular functions and, in some cases, whole organisms' activities, both relative to their metagenomic abundances (Box 1). For example, combined metatranscriptomic and metagenomic sequencing of the healthy human gut has revealed that the biosynthesis of small molecules (such as tryptophan and other amino acids) tends to be under-expressed in this environment (as the DNA abundance consistently exceeds RNA abundance), presumably because these compounds are readily available from the host and thus their synthesis by microorganisms would be energetically unfavorable[25] (Figure 3).

Sporulation was also strongly inactivated, presumably because bacteria are growing under ideal conditions in the healthy human gut (Figure 3).

By contrast, genes associated with methanogenesis in the archaeal species *Methanobrevibacter smithii* were strongly over-represented (as the RNA abundance consistently exceeded DNA abundance) in the healthy gut metatranscriptome, indicating a heightened level of transcriptional activity relative to other gut microorganisms (Figure 3). Similarly, *tetA* (an antibiotic resistance determinant), *groEL* (a chaperone protein) and bacterial ribosomal genes were also strongly over-expressed (RNA abundance exceeded DNA abundance), which suggests that these functions are highly active in the human gut.Interestingly, transcription of genes encoding bacterial ribosomal proteins and *groEL* were highly variable across individuals, which is consistent with a pattern of subject-specific transcriptional regulation (Figure 3). Such inferences would not be possible if microbial community RNA or DNA sequence data were considered in isolation.

Outside of the human gut, combined meta'omic profiling has been applied to the subgingival plaque of individuals with periodontitis, revealing an unexpected degree of transcriptional reorganization among canonically non-pathogenic bacteria, including the over-expression of putative virulence factors[47]. These findings suggest a role for metatranscriptomics in identifying bacterial species that influence disease through mechanisms that do not involve overgrowth and that would otherwise be missed by metagenomics- or culture-based assays. At the same time, over-transcription of putative virulence factors by canonically non-pathogenic bacteria could suggest that these factors are engaged in other non-pathogenic processes (and hence their annotations are incomplete) or that these bacteria should be reclassified as opportunistic pathogens.

In an environmental context, meta'omic sequencing of waters contaminated by the Deepwater Horizon oil spill revealed enrichment for species and pathways involved in the degradation of complex hydrocarbons[48]. However, RNA data revealed that only those degradation pathways targeting simpler, aliphatic hydrocarbons were highly expressed, while pathways targeting more complex, aromatic compounds (such as benzene) remained largely inactivated. This suggests that combined meta'omic sequencing could play an important part in the design of bioremediation strategies, where it would be necessary to ensure that degradation pathways are both present and active in a microbial community.

Combining metagenomics with metatranscriptomics can also reveal changes in functional activity in response to perturbations, such as the changes in gene expression in the human microbiota in response to dietary[49] and xenobiotic[50] stimuli. For example, introducing a consortium of bacteria into the human or mouse gut via a fermented milk product (FMP) had minimal downstream effects on the composition of the native gut microbiota[49]. However, metatranscriptomics analysis revealed significant changes in microbial gene expression following introduction of the FMP, particularly in pathways related to carbohydrate metabolism; such changes would go undetected in a metagenomics-only approach. Finally, while recent surveys of the human gut microbiome have revealed a remarkable degree of conservation in the functional potential across individuals[7,9], metatranscriptomes seem to be

more personalized[25], which is indicative of possible "fine-tuning" of microbial gene expression in individuals (Figure 3).

Taken together, these studies suggest an ecological model in which a core metagenome (with constant functional potential) exists for a given environment, with functional elements that are conserved despite possible variations in the taxa that encode them, and where the functional activity is regulated via changes in gene expression. The temporal dynamics of this variation remain an open question, which could be answered by meta'omic sequencing in a longitudinal format[51].

Measuring functional activity with metaproteomics

Genes and transcripts are useful for the functional characterization of the activity of microorganisms because they are proxies for protein expression. However, measuring protein abundance provides a more direct measure of the functional activity of a cell or community. Protein abundance can be determined in a high-throughput manner using next-generation proteomics[52] (metaproteomics in the microbiome context). Proteomic methods rely on mass spectrometry-based shotgun quantification of peptide mass and abundance. Briefly, the fragmentation pattern of a peptide reveals both its amino acid sequence and any post-translational modifications (PTMs), such as phosphorylation. Peptides are then associated with full-length proteins by sequence homology-based searches against reference databases, similarly to the mapping of short nucleotide reads in metagenomic and metatranscriptomic profiling (Box 1).

Single-organism proteomics has suggested that a substantial fraction of biological regulation occurs at the level of protein expression and degradation[53,54]. This observation has naturally motivated the application of proteomic methods to study functional activity and regulatory phenomena in microbial communities. In the first comprehensive characterization of the healthy human gut metaproteome, over 50% of total microbial proteins were involved in housekeeping functions, including translation and energy production[55]. Comparative metaproteomics of the gut microbiome in subjects with Crohn's disease and healthy controls revealed significant changes in protein abundance for more than 100 protein families[56], including the depletion of proteins involved in short-chain fatty acid (SCFA) production among patients with Crohn's disease. Depletion of these compounds, which are proposed to play a role in reducing inflammation and promoting colonic health[57,58], may contribute to the pro-inflammatory state in these patients. Importantly, host proteins constitute up to one-third of the metaproteome from human stool samples[55], which allows the integrated analysis of host and microbial functions. For example, patients with Crohn's disease displayed lower levels of proteins involved in the maintenance of epithelial integrity and function, which is consistent with histological changes observed in these patients (such as epithelial barrier defects)[56].

Additional advantages of a metaproteomic approach are evident in analyses of microbiomes outside of the human body. For example, metaproteomics has been applied to monitor changes in biofilm formation in environmental communities associated with increased temperature, demonstrating an increased abundance of proteins involved in amino acid metabolism[59]. This technique has also been applied to assess the adaptation of marine

bacteria to oligotrophic (nutrient-depleted) environments[60], and identified an enrichment in peptides from three major marine-associated lineages: SAR11, *Prochlorococcus*, and *Synechococcus*. SAR11 peptides corresponding to nutrient capture were particularly enriched in these samples, including periplasmic phosphate- and amino acid-binding proteins. In addition, metaproteomic methods have been applied to specifically profile membrane-bound proteins in marine environments[61]. This work revealed a gradient of microbial transport functions in samples drawn from coastal versus open ocean sites: coastal communities were more enriched for TonB-dependent transporters [which bind and transport siderophores, vitamin B(12), nickel complexes, and carbohydrates], whereas open-ocean communities were more enriched for porins and permeases. These data suggest that expression of these transporters may provide microorganisms with a selective advantage in oligotrophic environments. Furthermore, this work highlights another advantage of metaproteomic profiling, which is the ability to target particular families of proteins based on their biophysical properties (by using upstream experimental enrichment strategies).

As a final example, metaproteomics analysis has been applied to a wastewater-associated microbial community sampled at multiple time points following exposure to cadmium[62]. Post-exposure changes in protein expression were grouped into three categories: functions that changed quickly following exposure but then returned to baseline (termed short-term resistance, which included the up-regulation of ATPases); changes that first occurred late in the time-course but were then maintained (termed long-term adaption, which included the up-regulation of secretory membrane proteins); and changes that occurred rapidly and were then maintained throughout the time-course analysis (termed sustained tolerance, which included the reconfiguration of metabolism). Like sequencing-based techniques, metaproteomic analysis can thus be combined with longitudinal sampling to investigate temporal variation in the functional activity of microbial communities. Furthermore, metaproteomics can reveal changes in microbial functional activity on short timescales, including changes that precede or occur in the absence of changes in community composition (such as the short-term resistance changes discussed above).

Measuring functional activity w€ith metabolomics

Metabolomics refers to the detection of metabolites and other small molecules in microbial communities. Notably, metabolomics as discussed here refers to direct, experimental quantification of metabolite abundances and not to predictions based on genomic composition (such as the identification of enzymes or reconstruction of pathways). Metabolomics relies on chromatography techniques (such as HPLC) to separate compounds followed by their identification and quantification based on mass spectrometry[63]. Metabolomics is therefore methodologically more similar to metaproteomics than to DNA or RNA sequencing. However, unlike metaproteomic profiling (which currently requires substantial biomass and unique sample preparations[55]), metabolomics methods are compatible with the sample biomass and preparations typical of meta'omic sequencing experiments.

While metabolomics shares several challenges with other meta'omic methods, including a broad catalog of potential features occurring with high dynamic range, it is further

complicated by the non-uniformity of the features (molecules) that are profiled[64]. For example, while all transcripts belong to the same class of biomolecules (RNAs), metabolites range from small, hydrophilic carbohydrates (such as glucose) to large, hydrophobic lipids (such as triacylglycerides), to complex natural compounds (such as antibiotics). Nevertheless, given that interactions between microorganisms or between microorganisms and their hosts are often mediated at the level of shared metabolite pools or metabolite exchanges, metabolomics remains a crucial tool for understanding the functional activity of microbial communities.

SCFAs provide an illustrative example of the importance of metabolites in microbiota-host interactions. As discussed above, these small molecules are excreted by bacteria in the gut and promote colonic health. For example, the SCFA butyrate is produced by *Bifidobacteria* in the gut and has been shown to have anti-tumorigenic effects[65]. In the earlier example from Crohn's disease, changes in SCFA profiles were inferred based on expression of the proteins involved in their production[56]. As small molecules, SCFAs can also be directly profiled by metabolomics methods. Indeed, one such study revealed significant differences in the SCFA profiles of healthy subjects versus those of patients with colorectal cancer, including a marked depletion for butyrate[66].

Metabolomics experiments have also revealed a number of bacterial metabolic products with negative effects on human health. For example, the presence of trimethylamine N-oxide (TMAO) in blood plasma has been linked to cardiovascular disease (CVD)[67]. The same study demonstrated that the gut microbiota was involved in the generation of TMAO from phosphatidylcholine, a dietary lipid. L-carnitine (an abundant compound in red meat) has also been linked to CVD: mice fed a diet rich in L-carnitine experienced changes in gut microbiome composition leading to increased TMAO production and CVD; these effects were reduced when the gut microbiota was suppressed with antibiotics[68]. Notably, the genus *Prevotella* was among the clades that expanded significantly in the carnitine-fed mice and was associated with higher blood plasma TMAO levels in humans.

Metabolomics has also revealed associations between the human microbiome and xenobiotic compounds (such as pharmaceutical drugs). For example, the efficacy of statins in lowering cholesterol levels was found to be inversely correlated with plasma levels of bacteria-derived bile acids (including lithocholic acid and its derivatives)[69]. This negative correlation could arise from competitive binding of bile acids and statins to shared transporter proteins, or through bile acid-mediated stabilization of cholesterol in blood plasma. Conversely, subjects with higher levels of coprostanol (a byproduct of bacterial metabolism of cholesterol) were predicted to respond more favorably to statin therapy. In addition, recent findings indicate that the gut microbiota can directly metabolize xenobiotics ingested by their hosts to inactive forms, and that byproducts of this xenobiotic degradation could also potentially exhibit unexpected, harmful activities. For example, digoxin, which is a pharmaceutical drug prescribed in the treatment of cardiac arrhythmias, is reduced by some *Eggerthella lenta* strains, thus reducing treatment efficacy[70]. Notably, this activity was only observed for strains encoding a pair of cardiac glycoside reductase enzymes (*cgr1* and *cgr2*). Arginine was shown to inhibit the expression of these enzymes, leading to increased digoxin levels both *in vitro* and in a mouse model. Hence, metabolic profiling in this study revealed

a xenobiotic metabolite (digoxin) that is implicated in an adverse host-microbe interaction and a second, dietary metabolite (arginine) that could mitigate the adverse effect.

One limitation of community metabolomics analysis is that, thus far, it has heavily focused on human-associated microbiomes. This is due in part to the fact that *ex vivo* environmental backgrounds (such as soil) possess properties that make them less amenable to standard metabolomics practices (such as high salt concentrations). However, a recent study proposed a metabolomics protocol that was specifically adapted for environmental samples through a stable isotope labeling step[71]. Samples from the environmental community of interest (in this case an acid mine drainage site) were cultured with ^{15}N-labeled ammonium sulfate as a nitrogen source. Compounds isolated by mass/charge ratio could then be more precisely identified based on their ^{15}N content. This study identified taurine as an important metabolite in the acid mine drainage environment under investigation, possibly due to its role in adaptation to osmotic stress, and metagenomics analysis revealed that taurine was most likely metabolized by the fungus *Acidomyces richmondensis*. Given the immense potential of metabolomic data for clarifying complex interactions in microbial communities, we expect that the development of improved methods in this area will remain a topic of great interest in the near future.

Integrating multi'omic data

The studies and methods introduced in the preceding sections highlighted many advantages of collecting additional multi'omics data types beyond DNA sequences in the characterization of microbial communities (Figure 4). RNA, protein, and metabolites all provide pictures of the functional activity of a community and this often differed markedly from the functional potential one would infer from DNA sequence alone (Figure 3). However, simply collecting more data types is not enough: making full use of multi'omic data requires a careful data integration strategy. Such a strategy begins at the experimental design phase, where one must trade-off between the number of communities sampled and the number of multi'omics assays performed per sample (Figure 1). One must also carefully consider the choice of analysis methods necessary for integrating multi'omics data types, some of which will be general to all studies, while others will depend on the particular questions under investigation.

Notably, many of the studies introduced above as successful applications of particular multi'omic data types also collected metagenomic sequencing data, making many of their analyses and results inherently integrative. For example, studies using RNA sequencing or proteomics to measure the functional activity of a particular gene will tend to normalize these data against the metagenomic abundance of that same gene (gene copy number); by not doing so, functional activity measurements would be confounded with the functional potential of the community (Figure 4). For example, in the absence of gene copy number data, the failure to detect a particular transcript could indicate that the gene encoding the transcript was either not expressed or simply not present. Combining DNA and RNA data allows these possibilities to be disentangled.

In the example of marine microbial gene expression following the Deepwater Horizon oil spill, the authors combined DNA and RNA data to demonstrate that pathways for degrading complex aromatic hydrocarbons were not expressed despite being encoded by the community[48]. Conversely, normalizing RNA data with DNA data can reveal genes, functions, or clades that are overrepresented in the transcriptional pool; this was the case for *M. smithii* in the human gut[25], whose methanogenesis pathways were strongly over-represented in the transcript pool relative to their metagenomic abundance (Figure 3). Such insights would not have been possible considering DNA or RNA data in isolation.

Another focus of data integration techniques is to combine multiple (potentially noisy and heterogeneous) signals to build support for specific hypotheses. The intuition here is simple: if independent lines of evidence arrive at the same conclusion, then our confidence in that conclusion grows (Figure 4). Such techniques have been widely applied to multi'omics data generated from model organisms in order to assign putative functions to genes[72,73] and to predict functional relationships between pairs of genes[74,75], including the reconstruction of physical interactions[76] and the identification of genes involved in the same pathway[77].

While these techniques have been developed for use in model organisms, they are fully applicable to microbial communities, although they have not yet seen wide application. This is in some ways understandable due to the complexity of microbial communities and the limited data that is available for analysis. Nevertheless, clear advantages of integrating independent multi'omics data for microbial studies are evident in the studies introduced in the preceding sections. For example, the two microbiological studies of human colon disease (Crohn's disease and colorectal cancer) employed two different multi'omics methods (proteomics and metabolomics) to show that disease conditions were associated with shifts in microbial metabolism of short chain fatty acids (SFCAs)[56,66]. Assuming one knew nothing else about this system, integrating data from these two experiments would lend support to a hypothesis that SCFAs are linked to human colonic health. As our ability to generate new data describing microbial communities – and the amount of data thus generated – continues to grow quickly, such data will remain underutilized unless integrative techniques are used to combine data within and across studies.

Lastly, integrative techniques are critical for finding associations between distinct data types and for filling mechanistic gaps. No single assay is capable of describing a microbial community in complete mechanistic detail, and a deeper description of the community can only be formed by considering multiple data types simultaneously (Figure 4). Consider the example of *E. lenta* and its interaction with the human pharmaceutical drug digoxin[70]. In that case, metabolomic data was used to demonstrate that *E. lenta* reduced digoxin to an inactive form and that this interaction was suppressed by arginine. Strain-level profiling revealed this to be a strain-specific effect, as only a subset of *E. lenta* strains (those encoding the *cgr* operon) interacted in this manner. Metatrancriptomics data revealed a transcriptional regulatory mechanism underlying the interaction: the *cgr* operon was up-regulated in the presence of digoxin, but the up-regulation was dampened in the presence of arginine. Achieving this level of detail in a descriptive model of a biological phenomenon depends critically on the integration of multiple data types. Notably, identifying such mechanisms

from high-dimensional multi'omics data requires special statistical techniques and considerations (Box 2).

In addition to descriptive modeling, a further goal of data integration is the construction of predictive models. One such area of focus is the construction of metabolic networks (reviewed in detail in [78]). Briefly, these methods involve the creation of a network in which metabolites are linked as reactants and products of enzymatic reactions. Constraint-based modeling (such as flux balance analysis) is then applied to the network to predict metabolic phenotypes under different growth conditions[79], possibly incorporating other multi'omic measurements, such as enzyme expression levels[80]. There is a growing interest in transitioning these methods from single organisms to communities of two or more species, which can be accomplished by modeling each species as a compartment in a larger network model that also incorporates exchanges of metabolites between the species. For example, such models enable predictions of waste products exported by one organism that can be imported as a resource by a second organism[81,82]. These techniques are also amenable to predicting metabolic interactions between microorganisms and their hosts[83], including cases where microorganisms produce critical nutrients for their host using reactions that are absent or defective in the host's own genomically-encoded metabolic network.

Recent years have seen a drive to expand metabolic network models to larger microbial communities, including human microbiomes. For example, the analysis of the metabolic relationships between distinct layers of an oral biofilm revealed that adjacent biofilm layers tended to be globally more metabolically similar than would be expected in a random ordering of the layers[84]. At the same time, adjacent layers were proposed to complement one another by contributing distinct and potentially synergistic metabolic modules to the biofilm. Conclusions drawn from such models will need to be further refined in the future by integrating additional multi'omics data types, such as by validating proposed metabolic exchange relationships using metabolomic profiling.

Summary and outlook

In order to take the next steps forward in understanding the basic biology of microbial communities, richer multi'omic studies will be necessary for both human-associated and environmental microbiomes. This goal can be partially accomplished by adapting current sequencing techniques to probe underappreciated aspects of microbial community behavior, such as strain-level phenomena, temporal dynamics, and functional activity (Figure 1; Figure 4). However, to understand more completely the nature and mechanisms of microbial community function and environmental interactions will require the development and application of alternative, high-throughput molecular biological screens. Success in this area will not be possible without the widespread adoption of integrative methods for managing and exploring such data. These include basic statistical considerations, such as methods for normalizing functional activity measurements against metagenomic potential, as well as continued application and development of supervised and unsupervised approaches for identifying patterns in large multi'omic data collections (Box 2).

In the context of model organisms, data integration methods have been invaluable for combining results from an ever-increasing number of independent studies and assays. These efforts have markedly improved both the coverage and accuracy of functional annotations assigned to biomolecules from these species. The application of these techniques to microbial communities is especially relevant given that large fractions of the biomolecules they contain have no assigned functions[6,85], suggesting the need for efficient but comprehensive efforts to characterize this basic "parts list." Notably, this is to a large degree true of microbial isolates as well[86,87], raising the possibility that many gene products may actually be easier to characterize in communities, as they might be functional in this context but inactive in laboratory monocultures. Likewise, few steps have yet been taken towards richer predictive modeling of microbial communities, incorporating regulatory relationships, ecology, or inter-organismal signaling in addition to metabolism. Such methods have been benchmarked in the context of single organisms and macroscopic ecological communities, and so their application to microbial communities is a natural next step for the field. Integrating information into models of community systems biology and, in turn, systems ecology will require both extensive multi'omic data collection and the development of controlled, perturbable model systems that accurately reflect "wild" microbial communities *in vitro*.

Finally, extensive efforts are already underway to translate the growing understanding of human-associated microbial communities into clinical biomarkers and treatments. Some areas, such as the treatment of *C. difficile* infection, have been tremendously successful[88], even prior to the development of accompanying mechanistic or ecological explanations. However, others diseases, such as inflammatory bowel disease (IBD)[11-13], seem more complex and successful microbiota-based treatments may require a deep understanding of the complex mechanisms of host-microbiota interactions, which could be elucidated by integrating host multi'omic data (such as gene expression, epigenetics, SNPs, proteomics and metabolomics) with microbiome data (such as strain variation, gene expression, proteomics and metabolomics)[89]. Although the field of microbial community studies continues to grow rapidly, fueled in part by the power and efficiency of sequencing-based investigative tools, considerable work remains to be done in refining these tools and integrating them into rich study designs for understanding microbial community biology.

Acknowledgments

This work was funded in part by NIH grants R01HG005969 and U54DK102557 (CH and Ramnik J. Xavier); NSF grant DBI-1053486 (CH); Army Research Office grant W911NF-11-1-0473 (CH); and Danone Research grant PLF-5972-GD (Wendy S. Garrett).

References

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998; 5:R245–9. [PubMed: 9818143]
2. Tyson GW, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004; 428:37–43. [PubMed: 14961025]
3. Venter JC, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004; 304:66–74. **One of the first large-scale environmental metagenomic sequencing projects;**

presented profiles of taxonomic composition and function from geographically diverse marine microbial communities. [PubMed: 15001713]

4. Rondon MR, et al. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol. 2000; 66:2541–7. [PubMed: 10831436]
5. Kembel SW, et al. Architectural design influences the diversity and structure of the built environment microbiome. ISME J. 2012; 6:1469–79. [PubMed: 22278670]
6. Qin J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010; 464:59–65. **The first large-scale exploration of the human microbiome using metagenomic sequencing; profiled the gene content of 124 European gut microbiomes, highlighting orders of magnitude more microbial genes than possessed by the human host, a large fraction of which were shared across individuals.** [PubMed: 20203603]
7. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012; 486:207–14. **The largest and most complete survey of the healthy human microbiome to date; sampled up to 18 distinct body sites in >200 individuals at multiple time points, enabling quantitative assessment of microbiome structure and stability across environments, individuals, and time.** [PubMed: 22699609]
8. Turnbaugh PJ, et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006; 444:1027–31. [PubMed: 17183312]
9. Turnbaugh PJ, et al. A core gut microbiome in obese and lean twins. Nature. 2009; 457:480–4. [PubMed: 19043404]
10. Ley RE. Obesity and the human microbiome. Curr Opin Gastroenterol. 2010; 26:5–11. [PubMed: 19901833]
11. Manichanh C, et al. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006; 55:205–11. [PubMed: 16188921]
12. Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012; 13:R79. **A metagenomic assessment of perturbations of the human gut microbiome in inflammatory bowel disease; revealed that changes in functional composition were more pronounced than changes in community membership.** [PubMed: 23013615]
13. Berry D, Reinisch W. Intestinal microbiota: a source of novel biomarkers in inflammatory bowel diseases? Best Pract Res Clin Gastroenterol. 2013; 27:47–58. [PubMed: 23768552]
14. Marchesi JR, et al. Towards the human colorectal cancer microbiome. PLoS ONE. 2011; 6:e20447. [PubMed: 21647227]
15. Kostic AD, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012; 22:292–8. [PubMed: 22009990]
16. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. Nat Rev Microbiol. 2012; 10:575–82. [PubMed: 22728587]
17. Eckburg PB, et al. Diversity of the human intestinal microbial flora. Science. 2005; 308:1635–8. **An early metagenomic survey of the human gut microbiome that considered both stool and mucosal samples; revealed many previously uncultured taxa along with strong inter-subject and inter-site differences.** [PubMed: 15831718]
18. Karch H, Tarr PI, Bielaszewska M. Enterohaemorrhagic Escherichia coli in human medicine. Int J Med Microbiol. 2005; 295:405–18. [PubMed: 16238016]
19. Eren AM, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol. 2013; 4 **This paper presents a computational approach for improving taxonomic resolution in surveys of microbial communities based on 16S sequencing.**
20. Eren AM, et al. Exploring the diversity of Gardnerella vaginalis in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. PLoS ONE. 2011; 6:e26732. [PubMed: 22046340]
21. McLellan SL, et al. Sewage reflects the distribution of human faecal Lachnospiraceae. Environ Microbiol. 2013; 15:2213–27. [PubMed: 23438335]

22. Faith JJ, et al. The long-term stability of the human gut microbiota. Science. 2013; 341 1237439. **This paper presents a sequencing method for improving taxonomic resolution in surveys of microbial communities based on 16S sequencing; the method was applied to quantify the stability of the human gut microbiome over a 5-year period.**
23. Segata N, et al. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012; 9:811–4. [PubMed: 22688413]
24. Scher JU, et al. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. eLife. 2013; 2:e01202. [PubMed: 24192039]
25. Franzosa EA, et al. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014; 111:E2329–38. [PubMed: 24843156]
26. Schloissnig S, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013; 493:45–50. [PubMed: 23222524]
27. Wrighton KC, et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. Science. 2012; 337:1661–5. [PubMed: 23019650]
28. Albertsen M, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013; 31:533–8. [PubMed: 23707974]
29. Castelle CJ, et al. Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nat Commun. 2013; 4:2120. [PubMed: 23979677]
30. Di Rienzi SC, et al. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. eLife. 2013; 2:e01102. [PubMed: 24137540]
31. Lax S, et al. Longitudinal analysis of microbial interaction between humans and the indoor environment. Science. 2014; 345:1048–52. [PubMed: 25170151]
32. Caporaso JG, et al. Moving pictures of the human microbiome. Genome Biol. 2011; 12:R50. [PubMed: 21624126]
33. Gronlund MM, Lehtonen OP, Eerola E, Kero P. Fecal microflora in healthy infants born by different methods of delivery: permanent changes in intestinal flora after cesarean delivery. J Pediatr Gastroenterol Nutr. 1999; 28:19–25. [PubMed: 9890463]
34. Sharon I, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013; 23:111–20. [PubMed: 22936250]
35. Kong HH, et al. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. Genome Res. 2012; 22:850–9. [PubMed: 22310478]
36. Khoruts A, Dicksved J, Jansson JK, Sadowsky MJ. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent Clostridium difficile-associated diarrhea. J Clin Gastroenterol. 2010; 44:354–60. [PubMed: 20048681]
37. Turnbaugh PJ, et al. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med. 2009; 1:6ra14.
38. Spencer MD, et al. Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. Gastroenterology. 2011; 140:976–86. [PubMed: 21129376]
39. Rodriguez-Brito B, et al. Viral and microbial community dynamics in four aquatic environments. ISME J. 2010; 4:739–51. [PubMed: 20147985]
40. Giannoukos G, et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 2012; 13:R23. [PubMed: 22455878]
41. Poretsky RS, et al. Analysis of microbial gene transcripts in environmental samples. Appl Environ Microbiol. 2005; 71:4121–6. [PubMed: 16000831]
42. Gilbert JA, et al. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS ONE. 2008; 3:e3042. [PubMed: 18725995]
43. Frias-Lopez J, et al. Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci U S A. 2008; 105:3805–10. [PubMed: 18316740]
44. Zhang T, et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 2006; 4:e3. [PubMed: 16336043]

45. Culley AI, Lang AS, Suttle CA. Metagenomic analysis of coastal RNA virus communities. Science. 2006; 312:1795–8. [PubMed: 16794078]
46. Willner D, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS ONE. 2009; 4:e7370. [PubMed: 19816605]
47. Duran-Pinedo AE, et al. Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. ISME J. 2014
48. Mason OU, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J. 2012; 6:1715–27. [PubMed: 22717885]
49. McNulty NP, et al. The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. Sci Transl Med. 2011; 3:106ra106.
50. Maurice CF, Haiser HJ, Turnbaugh PJ. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell. 2013; 152:39–50. [PubMed: 23332745]
51. Gilbert JA, et al. The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. PLoS ONE. 2010; 5:e15545. [PubMed: 21124740]
52. Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet. 2013; 14:35–48. [PubMed: 23207911]
53. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. Mol Cell Biol. 1999; 19:1720–30. [PubMed: 10022859]
54. Schwanhausser B, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–42. [PubMed: 21593866]
55. Verberkmoes NC, et al. Shotgun metaproteomics of the human distal gut microbiota. ISME J. 2009; 3:179–89. [PubMed: 18971961]
56. Erickson AR, et al. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS ONE. 2012; 7:e49138. [PubMed: 23209564]
57. Smith PM, et al. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science. 2013; 341:569–73. **This demonstrated that short-chain fatty acids, a common class of microbial metabolites, play an important role in coadaptation between the gut microbiome and host immune system.** [PubMed: 23828891]
58. Furusawa Y, et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. Nature. 2013; 504:446–50. [PubMed: 24226770]
59. Mosier AC, et al. Elevated temperature alters proteomic responses of individual organisms within a biofilm community. ISME J. 2015; 9:180–94. [PubMed: 25050524]
60. Sowell SM, et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. 2009; 3:93–105. [PubMed: 18769456]
61. Morris RM, et al. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J. 2010; 4:673–85. [PubMed: 20164862]
62. Lacerda CM, Choe LH, Reardon KF. Metaproteomic analysis of a bacterial community response to cadmium exposure. J Proteome Res. 2007; 6:1145–52. [PubMed: 17284062]
63. Turnbaugh PJ, Gordon JI. An invitation to the marriage of metagenomics and metabolomics. Cell. 2008; 134:708–13. [PubMed: 18775300]
64. Tang J. Microbial metabolomics. Curr Genomics. 2011; 12:391–403. [PubMed: 22379393]
65. Williams EA, Coxhead JM, Mathers JC. Anti-cancer effects of butyrate: use of micro-array technology to investigate mechanisms. P Nutr Soc. 2003; 62:107–15.
66. Weir TL, et al. Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. PLoS ONE. 2013; 8:e70803. [PubMed: 23940645]
67. Wang Z, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. 2011; 472:57–63. **Integrated multi'omic analysis combined with experimental work in mice demonstrated a functional link between metabolism of dietary compounds by the gut microbiome and the development of cardiovascular disease.** [PubMed: 21475195]
68. Koeth RA, et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med. 2013; 19:576–85. [PubMed: 23563705]

69. Kaddurah-Daouk R, et al. Enteric microbiome metabolites correlate with response to simvastatin treatment. PLoS ONE. 2011; 6:e25482. [PubMed: 22022402]
70. Haiser HJ, et al. Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science. 2013; 341:295–8. **Integrated multi'omic analysis identified an operon in a member of the human gut microbiome community involved in degradation (and hence, loss of efficacy) of the cardiac drug digoxin.** [PubMed: 23869020]
71. Mosier AC, et al. Metabolites associated with adaptation of microorganisms to an acidophilic, metal-rich environment identified by stable-isotope-enabled metabolomics. mBio. 2013; 4:e00484–12. [PubMed: 23481603]
72. Karaoz U, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci U S A. 2004; 101:2888–93. [PubMed: 14981259]
73. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci U S A. 2003; 100:8348–53. [PubMed: 12826619]
74. Lee I, Date SV, Adai AT, Marcotte EM. A probabilistic functional network of yeast genes. Science. 2004; 306:1555–8. [PubMed: 15567862]
75. Myers CL, et al. Discovery of biological networks from diverse functional genomic data. Genome Biol. 2005; 6:R114. [PubMed: 16420673]
76. Jansen R, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003; 302:449–53. [PubMed: 14564010]
77. Park CY, Hess DC, Huttenhower C, Troyanskaya OG. Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. PLoS Comput Biol. 2010; 6:e1001009. [PubMed: 21124865]
78. Thiele I, Palsson BO. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc. 2010; 5:93–121. [PubMed: 20057383]
79. Durot M, Bourguignon PY, Schachter V. Genome-scale models of bacterial metabolism: reconstruction and applications. FEMS Microbiol Rev. 2009; 33:164–90. [PubMed: 19067749]
80. Bordel S, Agren R, Nielsen J. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. PLoS Comput Biol. 2010; 6:e1000859. [PubMed: 20657658]
81. Stolyar S, et al. Metabolic modeling of a mutualistic microbial community. Mol Syst Biol. 2007; 3:92. [PubMed: 17353934]
82. Klitgord N, Segre D. Environments that induce synthetic microbial ecosystems. PLoS Comput Biol. 2010; 6:e1001002. [PubMed: 21124952]
83. Heinken A, Sahoo S, Fleming RM, Thiele I. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut microbes. 2013; 4:28–40. [PubMed: 23022739]
84. Mazumdar V, Amar S, Segre D. Metabolic proximity in the order of colonization of a microbial community. PLoS ONE. 2013; 8:e77617. [PubMed: 24204896]
85. Howe AC, et al. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci U S A. 2014; 111:4904–9. [PubMed: 24632729]
86. Roberts RJ, et al. COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. Nucleic Acids Res. 2011; 39:D11–4. [PubMed: 21097892]
87. Harrington ED, et al. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A. 2007; 104:13913–8. [PubMed: 17717083]
88. Gough E, Shaikh H, Manges AR. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent Clostridium difficile infection. Clin Infect Dis. 2011; 53:994–1002. [PubMed: 22002980]
89. Bhavsar AP, Guttman JA, Finlay BB. Manipulation of host-cell pathways by bacterial pathogens. Nature. 2007; 449:827–34. [PubMed: 17943119]
90. Riesenfeld CS, Schloss PD, Handelsman J. Metagenomics: genomic analysis of microbial communities. Annu Rev Genet. 2004; 38:525–52. [PubMed: 15568985]
91. Hamady M, Knight R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 2009; 19:1141–52. [PubMed: 19383763]

92. Segata N, et al. Computational meta'omics for microbial community studies. Mol Syst Biol. 2013; 9:666. **An in-depth review of computational methods in microbial community analysis.** [PubMed: 23670539]

93. McDonald D, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012; 6:610–8. [PubMed: 22134646]

94. Yilmaz P, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014; 42:D643–8. [PubMed: 24293649]

95. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ Microbiol. 2010; 12:1889–98. [PubMed: 20236171]

96. Knights D, Parfrey LW, Zaneveld J, Lozupone C, Knight R. Human-associated microbial signatures: examining their predictive value. Cell Host Microbe. 2011; 10:292–6. [PubMed: 22018228]

97. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 2007; 4:63–72. [PubMed: 17179938]

98. Sunagawa S, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013; 10:1196–9. [PubMed: 24141494]

99. Brady A, Salzberg S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods. 2011; 8:367. [PubMed: 21527926]

100. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15:R46. [PubMed: 24580807]

101. Kanehisa M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 2014; 42:D199–205. [PubMed: 24214961]

102. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997; 278:631–7. [PubMed: 9381173]

103. Powell S, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012; 40:D284–9. [PubMed: 22096231]

104. Punta M, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40:D290–301. [PubMed: 22127870]

105. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007; 23:1282–8. [PubMed: 17379688]

106. Langille MG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013; 31:814–21. [PubMed: 23975157]

107. Caspi R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014; 42:D459–71. [PubMed: 24225315]

108. Overbeek R, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005; 33:5691–702. [PubMed: 16214803]

109. Markowitz VM, et al. IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic Acids Res. 2012; 40:D123–9. [PubMed: 22086953]

110. Meyer F, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9:386. [PubMed: 18803844]

111. Konwar KM, Hanson NW, Page AP, Hallam SJ. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. BMC Bioinformatics. 2013; 14:202. [PubMed: 23800136]

112. Abubucker S, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012; 8:e1002358. [PubMed: 22719234]

113. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007; 62:142–60. [PubMed: 17892477]

114. Gianoulis TA, et al. Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci U S A. 2009; 106:1374–9. **An in-depth review of statistical**

procedures for identifying patterns in high-dimensional microbial community data. [PubMed: 19164758]

115. McHardy IH, et al. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome. 2013; 1:17. [PubMed: 24450808]
116. Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS Microbiol Rev. 2011; 35:343–59. [PubMed: 21039646]
117. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput Biol. 2009; 5:e1000352. [PubMed: 19360128]
118. Segata N, et al. Metagenomic biomarker discovery and explanation. Genome Biol. 2011; 12:R60. [PubMed: 21702898]
119. Tickle TL, Segata N, Waldron L, Weingart U, Huttenhower C. Two-stage microbial community experimental design. ISME J. 2013; 7:2330–9. [PubMed: 23949665]

Glossary

Contig	An assemblage of overlapping DNA or RNA reads from a high-throughput sequencing experiment
contigs capture larger	continuous sections of genomic (or transcript) material than those represented by individual reads
Flux Balance Analysis (FBA)	A computational method for (i) representing the steady-state metabolic network of an organism or community and (ii) evaluating its capacity to produce a set of target metabolites from a set of input metabolites
Horizontal gene transfer (HGT)	A process in which genetic material is transferred from one cell to the genome of another cell by a method other than normal reproduction (i.e. vertical transmission from a mother cell to daughter cell). HGT is also referred to as lateral gene transfer (LGT)
LEA-Seq	An alternative amplicon sequencing strategy designed to distinguish rare biological variation from sequencing errors, thus leading to more accurate profiling of low-abundance taxa in a community
Metagenomics	The application of high-throughput DNA sequencing to profile microbial community genomic composition in a culture-independent manner
Microbiome	The community, biomolecular repertoire, and ecology of microorganisms inhabiting a particular environment
Microbiota	The collection of microorganisms (of all types, bacteria, archaea, viruses, and eukaryotes) inhabiting a particular environment
Multi'omics	An experimental approach that combines two or more distinct high-throughput molecular biological (‘omics) assays. The resulting data are generally analyzed and combined by integrative methods

Oligotyping	A computational method for differentiating between closely related taxa in microbiomes profiled by amplicon sequencing
Read	A short DNA or RNA sequence derived from a high-throughput sequencing experiment. Reads are often described as "paired," which indicates that two sequences were derived from opposite ends of the same molecular DNA or RNA fragment
Single nucleotide polymorphism (SNP)	A position in a reference genome that occurs in more than one nucleotide state (A, C, G, T) among the members of a population
Sporulation	A stress response mechanism employed by (primarily Gram-positive) bacteria to survive periods of nutrient depletion

Box 1

Taxonomic and functional profiling of microbial communities

Sequence-based taxonomic profiling of a microbiome can be carried out using either amplicon (typically the 16S rRNA gene) or whole metagenome shotgun (WMS) sequencing (reviewed in [90-92]).

Amplicon sequencing

Amplicon sequences (reads) are either directly matched to reference taxa[93,94] or more commonly they are first grouped into clusters referred to as operational taxonomic units (OTUs) that share a fixed level of sequence identity (often 97%)[95,96]. In either case, individual reads or OTUs are then assigned to specific taxa based on sequence homology to a reference genomic sequence—a process referred to as "binning."

WMS sequencing

In this case, some or all shotgun reads are used to determine membership in a community, either by considering the reads individually or by first assembling them into contigs[97]. In one approach, short reads or contigs are profiled directly by comparison to a reference catalogue of microbial genes or genomes. In addition to quantifying species abundance, this approach can reveal strain-level variation (Figure 2), which manifests as small inconsistencies between the sample data and the reference catalogue (for example, a contig that is largely [but not entirely] explained by genes from a single species may contain a HGT event). Alternatively, individual reads can be mapped to a pre-computed catalog of clade-specific marker sequences (with[98] or without[23] pre-clustering); this approach tends to be more specific and is less computationally intensive than mapping reads to a comprehensive reference database. Finally, reads or contigs may be assigned to species based on agreement with models of genome composition[99] or by exact k-mer matching[100], thus enabling placement of reads or assembled contigs when corresponding reference genomes are not available (which is common for poorly characterized communities).

Functional profiling

This process usually begins by associating metagenomic and metatranscriptomic (collectively "meta'omic") sequence data with known gene families. This can be accomplished by directly mapping DNA or RNA reads to databases of gene sequences that have been clustered at the family level; such databases include KEGG Orthology[101], COG[102], NOG[103], Pfam[104], and UniRef[105]. Naturally, the number of reads that can be mapped in this manner depends on the completeness of the underlying reference database. Alternatively, reads can be assembled into contigs to determine putative protein-coding sequences (CDSs), which are then assigned to gene families following the same or similar methods used for annotating isolate microbial genomes. Both strategies yield profiles of the presence and absence of a gene family as well as the relative abundance of each family within a meta'omic sample. Amplicon sequencing is not amenable to this form of functional profiling as it typically only amplifies a single marker gene. Instead, functional profiles can be approximated for marker-based samples by

associating single gene sequences (such as the 16S rRNA gene) with annotated reference genomes; CDSs in those genomes are then likely to have been linked to the 16S rRNA or other marker gene copies in the original sample[106].

Pathway reconstruction

Functional profiles at the gene family-level may contain many thousands of features, so downstream analyses can be made more tractable by further performing per-organism or whole-community pathway reconstruction based on these genes. Although not specifically designed for microbial community analysis, species-specific pathway databases such as KEGG[101], MetaCyc[107], and SEED[108] can be useful for this purpose. Integrated bioinformatics pipelines such as IMG/M[109], MG-RAST[110], MetaPathways[111], and HUMAnN[112] have been developed to streamline the conversion of raw meta'omic sequencing data into more easily-interpreted profiles of microbial community function. Functional profiling methods have been reviewed further elsewhere[92].

Box 2

Statistical considerations in multi'omic data integration

Distinct multi'omic data types can be combined appropriately with exploratory, unsupervised approaches or using supervised statistical tests or classification.

Unsupervised approaches

Ordination is a common unsupervised analysis for microbial community taxonomic profiles that shows the largest patterns of variation in community composition (see Box 1). Common ordination methods include principle component analysis (PCA), principle coordinates analysis (PCoA), and non-metric multidimensional scaling (NMDS) (reviewed in an ecological context in reference [113]). Briefly, the goal of these methods is to project samples from a high-dimensional space (characterized by measurements of hundreds of metagenomic features, for example) into a two- or three-dimensional plot such that inter-sample distances in the plot best reflect true inter-sample distances. For example, PCoA was used to generate a broad overview of hundreds of samples collected during the HMP[7]. In that case it provided an efficient means for visualizing the ecological similarity of human skin and nasal samples relative to more diverse oral, urogenital, and gastrointestinal communities.

Multiple ordination methods identify dominant features of one set of measurements that co-vary with the dominant features of a second set of measurements. These methods are applicable when more than one multi'omic technique has been applied to the same set of samples. Canonical correlation analysis (CCA), for example, has been applied to marine microbiomes to identify broad relationships between pathway composition (such as energy-conversion strategies) and diverse environmental gradients (such as temperature)[114]. Another such method, Procrustes analysis, separately reduces two high-dimensional datasets to lower-dimensional spaces, as described above. The separate ordinations are then compared to see if they arrange the underlying samples in a similar manner, which would suggest that similarity in one space is associated with (and potentially influences) similarity in the second space. Procrustes analysis has been used to demonstrate strong coupling between microbial species composition and metabolite pools at two human gut mucosal surfaces, suggesting that mucosal microorganisms are producing these metabolites and/or dependent upon their production.[115]

Supervised approaches

In many cases supervised integration methods are more appropriate, as they reveal not only the largest patterns of variation in multi'omic data, but also their statistical significance and reproducibility. Such methods are central to metagenome-wide association studies (MWAS), which seek to link individual microbial features with other properties (such as disease status). MWAS shares many statistical considerations with expression quantitative trait locus (eQTL) analysis, which seeks to identify associations between a host's own genomic features and tissue-specific gene expression. Similarities between MWAS and eQTL analysis include complications from non non-normally distributed data and loss of power from performing many comparisons. Supervised

methods appropriate for microbiome data (reviewed in [116]) include standard machine learning techniques (such as random forests and support vector machines) as well as microbiome-specific tests (such as Metastats[117] and LEfSe[118]). Supervised integration methods have been crucial for identifying metagenomic biomarkers in a variety of human diseases, including obesity[8-10], inflammatory bowel disease[11-13], and cancer[14-16].

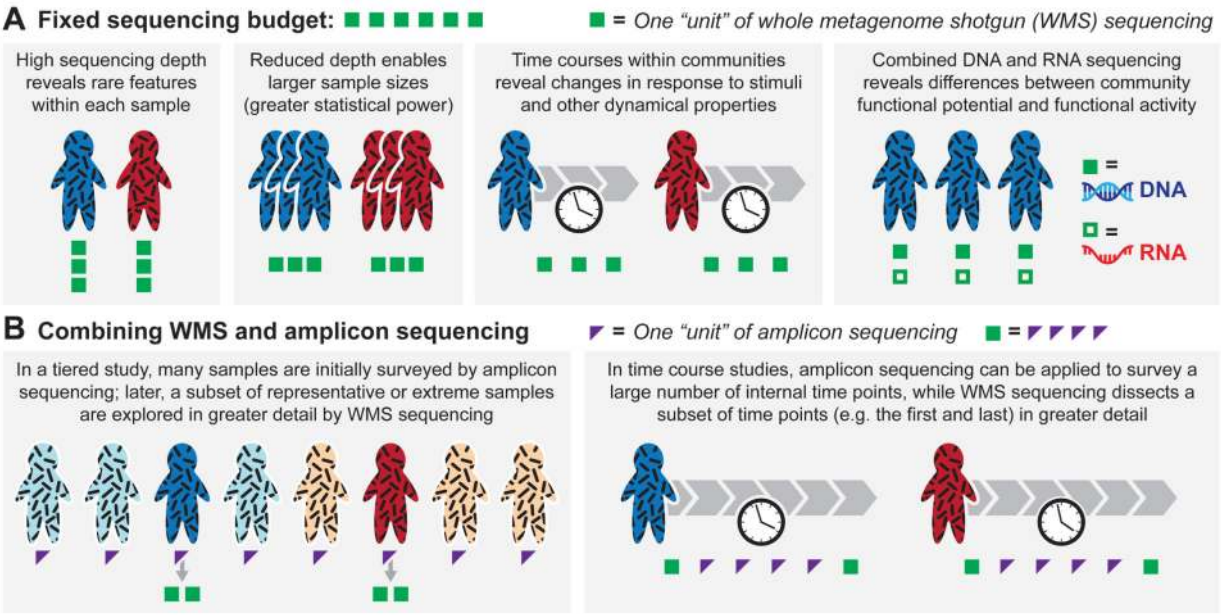

Figure 1. Optimizing experimental design

a. Whole metagenome shotgun (WMS) sequencing studies face trade-offs between the number of subjects considered, the number of samples per subject, and the sequencing depth per sample achievable with a fixed sequencing budget (here, six "units" of WMS sequencing). Greater sequencing depth facilitates identification of rare species and rare variants of abundant species (such as SNPs); considering more subjects improves statistical power in case-control studies; considering multiple samples per subject is critical for time course analysis; and combining DNA and RNA (meta'omic) sequencing reveals differences between the functional potential and the functional activity of the microbial communities present in different individuals. **b.** Combining the lower cost and decreased resolution of amplicon sequencing with the higher cost and increased resolution of WMS sequencing (here, one "unit" of WMS sequencing and four "units" of amplicon sequencing are considered to have equivalent costs) enables richer experimental designs. For example, two-stage study designs begin by surveying a large number of individuals using amplicon sequencing and then follow-up with a subset of samples using WMS sequencing (selected based on individuals that are representative of the group or those that represent the extreme cases within the group[119]). Similarly, time course studies can combine amplicon sequencing, which is used to survey a large number of time points, with WMS sequencing, which is applied to analyze a subset of time points (such as the first and last) in greater detail. Although depicted here in the context of sequencing-based assays in humans, these considerations are equally applicable to environmental samples and to a variety of high-throughput functional screens, including metaproteomics and metabolomics.

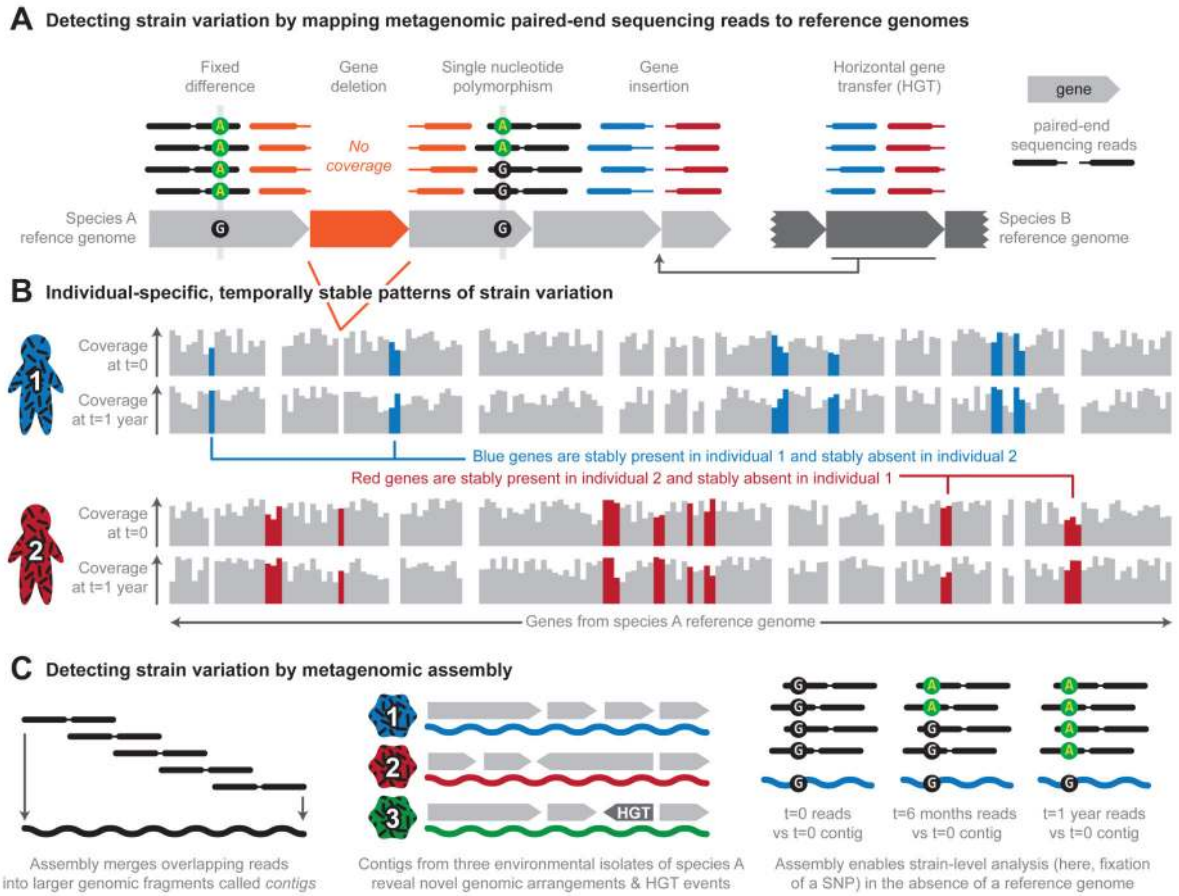

Figure 2. Profiling strain-level variation in microbial communities

a. Mapping paired-end sequencing reads to microbial reference genomes reveals not only the genomes that are present in a community, but also differences between the isolates of particular species and the reference isolate. In this example, most positions have 4x coverage, represented by 4 paired-end sequencing reads stacked above (mapped to) each position in the reference genomes. Gene deletion events can be detected with relatively low coverage of the reference genome; deleted genes (in orange) recruit no reads from the sample and are flanked by paired reads (orange paired reads). Higher coverage facilitates differentiating between sequencing error and true nucleotide-level strain variation. Such variation includes fixed differences (in which the sample is consistently different from the reference at some site) and single nucleotide polymorphisms (SNPs; in which a site occurs in two or more states in the sample). Paired reads that do not map together (red and blue reads) indicate additional structural variation, including the insertion of genomic material not found in the reference by mechanisms such as horizontal gene transfer (HGT). **b.** Assembling paired-end reads into larger genomic fragments, called contigs, facilitates detection of strain variation in the absence of a reference genome. For example, analyzing contigs from three environmental isolates of a microbial species can reveal novel genomic arrangements and HGT events. Metagenomic assembly also allows the comparison of reference contigs (in this case, t = 0) to paired-end reads obtained at different time points during temporal analysis (such as t = 6 months or t = 1 year), which enables the

identification of emerging SNPs. **c.** Mapping reads to reference genomes reveals patterns of gene presence and absence, which is a form of strain variation. Here, two individuals sampled at two time points (t = 0 and t =1 year) are distinguished by the presence and absence of genes in species A. The blue genes are stably present in individual 1 and stably absent in individual 2, whereas the red genes are stably present in individual 2 and stably absent in individual 1.

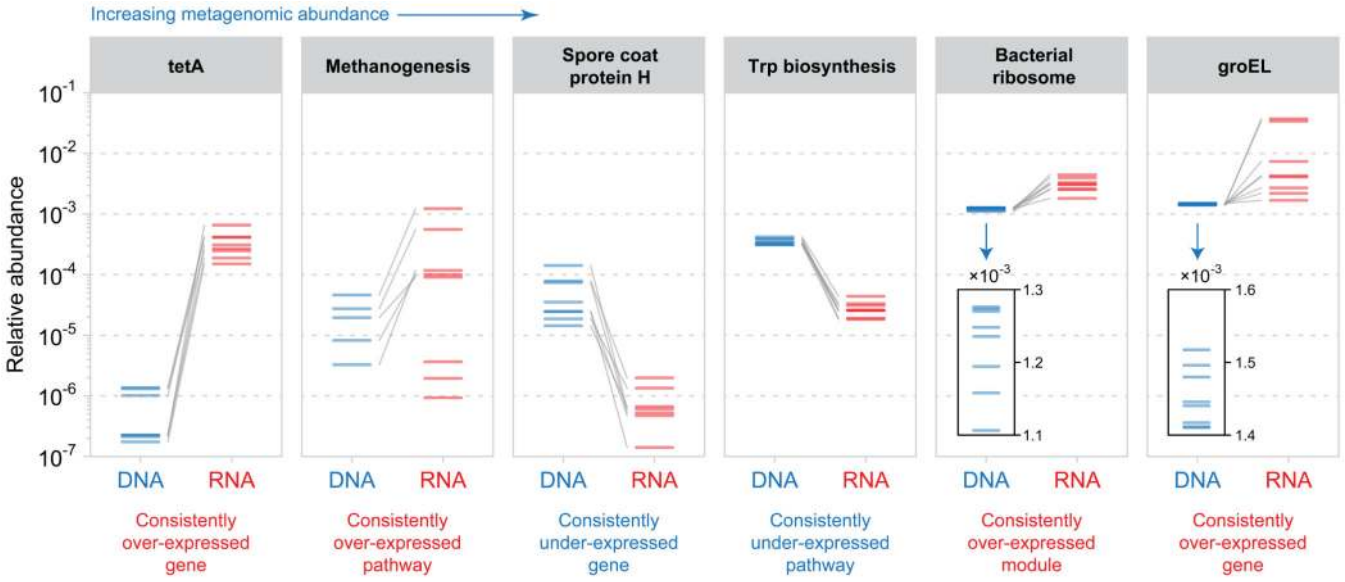

Figure 3. Relating the metatranscriptome and metagenome in the human gut

In this example, shotgun RNA and DNA sequence data from gut microbiome samples of 8 healthy individuals[25] were functionally profiled with HUMAnN[112]. Each panel illustrates a gene or functional module for which functional activity (expression level) deviated strongly from functional potential (metagenomic abundance). The median gene or transcript abundance is plotted for functions involving more than one gene; DNA and RNA values from the same individual are connected. tetA (an antibiotic resistance determinant), methanogenesis (an important metabolic pathway among gut archaeal species), the bacterial ribosome, and groEL (a bacterial chaperone protein) were strongly over-expressed, as their RNA abundance consistently exceeded their DNA abundance. Hence, on average, genes involved in these functions were producing many transcript copies, suggesting that they were highly active in the human gut (for example, that bacterial ribosomal subunits were being continuously synthesized). Conversely, spore coat protein H (a gene involved in response to nutrient starvation) and synthesis of the amino acid tryptophan were strongly under-expressed (DNA abundance consistently exceeded RNA abundance). Reduced transcription from these microbial functions suggests that they were down-regulated in the healthy human gut, likely due to the high bioavailability of nutrients (including tryptophan) derived from the host's diet. Transcription of bacterial ribosomal proteins and groEL were highly variable across individuals relative to their strong metagenomic conservation (see inset panels), which is consistent with a pattern of subject-specific transcriptional regulation. Such inferences would not be possible if microbial community RNA or DNA sequence data were considered in isolation.

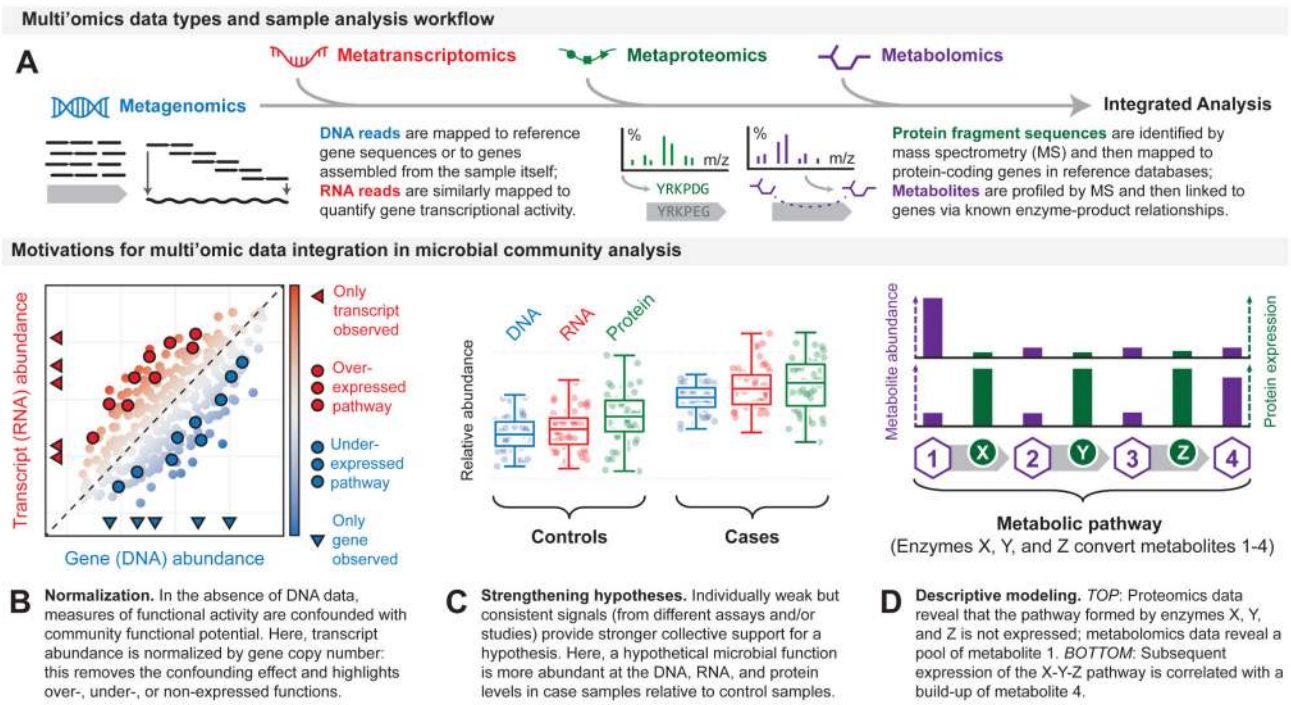

Figure 4. Integrating multi'omic data for deeper biological insights

a. To facilitate integrated analysis of a microbiome sample, distinct multi'omic data types are often associated with microbial genes or gene families that act as a shared point-of-reference. These genes may be taken from a reference database or directly assembled from the sample. Metagenomic, metatranscriptomic, and metaproteomic sequence data (such as paired-end reads or protein fragments identified by mass spectrometry) are then directly mapped to these genes based on sequence homology, which yields information about the copy numbers and activities of genes. Metabolites (identified by mass spectrometry) can be mapped to a subset of the genes by taking advantage of known relationships between enzyme-coding genes and their products, thus providing an additional, independent measure of gene activity. There are several motivations and advantages to perform multi'omic data integration. For example, in the absence of DNA data, measures of functional activity are confounded with community functional potential. Therefore, transcript abundance can be normalized by gene copy number; this removes the confounding effect and highlights over-, under-, or non-expressed functions (**part b**). Individually weak but consistent signals (from different assays and/or studies) provide stronger collective support for a hypothesis. Here, a hypothetical microbial function is more abundant at the DNA, RNA, and protein levels in case samples relative to control samples (**part c**). Data integration also enables descriptive modeling. For example, combining data from proteomics and metabolomics analyses can reveal whether a pathway formed by different enzymes (in this case X, Y, and Z, which metabolize substrates 1, 2 and 3, respectively) is inactive or active (**part d**).