MICROBIOME

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]

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 multiple 'omic' data types that better capture community functional activity, such as transcriptomics, proteomics and metabolomics; and combining multiple forms of 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.

Metagenomics
The application of high-throughput DNA sequencing to profile the genomic composition of a microbial community in a culture-independent manner.

Research into 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 the question of 'who is present in the community?'; and functional profiling, which answers the question of '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 ribosomal RNA (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) sequencing and amplicon sequencing have been applied to study diverse microbiomes, ranging from natural environments[2–4] to the built environment[5] and the human body[6,7]. For example, metagenomic profiling is used to monitor shifts in human microbiome composition and function associated with human diseases, including obesity[8–10], inflammatory bowel disease[11–13] and cancer[14–16]. Although 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 (for example, 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 behaviour of microbial communities. Addressing these issues requires new considerations at the experimental design phase, such as assessing the trade-offs between the number of subjects or environments considered (sample size (N)), the depth of sequencing per environment (greater depth facilitates strain-level analysis) and the number of time points considered for each subject or environment (FIG. 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 (FIG. 1).

Metagenomic sequencing faces a fundamental limitation in that it is unable to directly measure the functional

[1]Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts 02115, USA.
[2]The Broad Institute, Cambridge, Massachusetts 02142, USA.
[3]Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA.
Correspondence to C.H.
e-mail: chuttenh@hsph.harvard.edu

doi:10.1038/nrmicro3451
Published online 27 April 2015

Box 1 | Taxonomic and functional profiling of microbial communities

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

Amplicon sequencing

Amplicon sequences (reads) are either directly matched to reference taxa[93,94] or, more commonly, are first grouped into clusters that are referred to as operational taxonomic units (OTUs) and 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 on the basis of 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 (FIG. 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 horizontal gene transfer event). Alternatively, individual reads can be mapped to a pre-computed catalogue of clade-specific marker sequences (with[98] or without[23] pre-clustering); this approach tends to be more specific and it 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 Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology[101], Clusters of Orthologous Groups (COGs)[102], Non-supervised Orthologous Groups (NOGs)[103], Pfam[104] and UniProt Reference (UniRef) Clusters[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 using the same or similar methods as those used for annotating microbial genomes from isolates. 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 because 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 these 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 carrying out 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 Integrated Microbial Genomes with Microbial Samples (IMG/M)[109], Metagenomics Rapid Annotation using Subsystem Technology (MG-RAST)[110], MetaPathways[111], and the Human Microbiome Project Unified Metabolic Analysis Network (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].

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 data on the levels of community RNA (transcriptomics), proteins (proteomics) and metabolites (metabolomics), preferably in an integrated framework. In this Review, we discuss these new directions in microbiome research and highlight examples of next-generation metagenomic and integrated multi-omic approaches that have led to more advanced hypotheses, mechanisms and models of microbial community evolution and function.

New approaches in taxonomic profiling

The most common restrictions of traditional metagenomic analysis are limited taxonomic resolution, which is usually restricted to the species level, and lack of temporal resolution. However, new strategies are emerging that allow the study of strain-level variation and the dynamic behaviour 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 results in a subset of *E. coli* strains becoming highly pathogenic, such as the well-known serotype O157:H7 (REF. 18). Therefore, strain-level profiling, in particular the profiling of gene content based on WMS or single-cell sequencing, is needed to identify such variation in uncultured or unknown organisms (FIG. 2).

Microbiomes
The community composition, biomolecular repertoire and ecology of microorganisms inhabiting particular environments.

Multi-omic
An experimental approach that combines two or more distinct high-throughput molecular biological (omics) assays, such as genomics, transcriptomics, proteomics and metabolomics. The resulting data are generally analysed and combined by integrative methods.

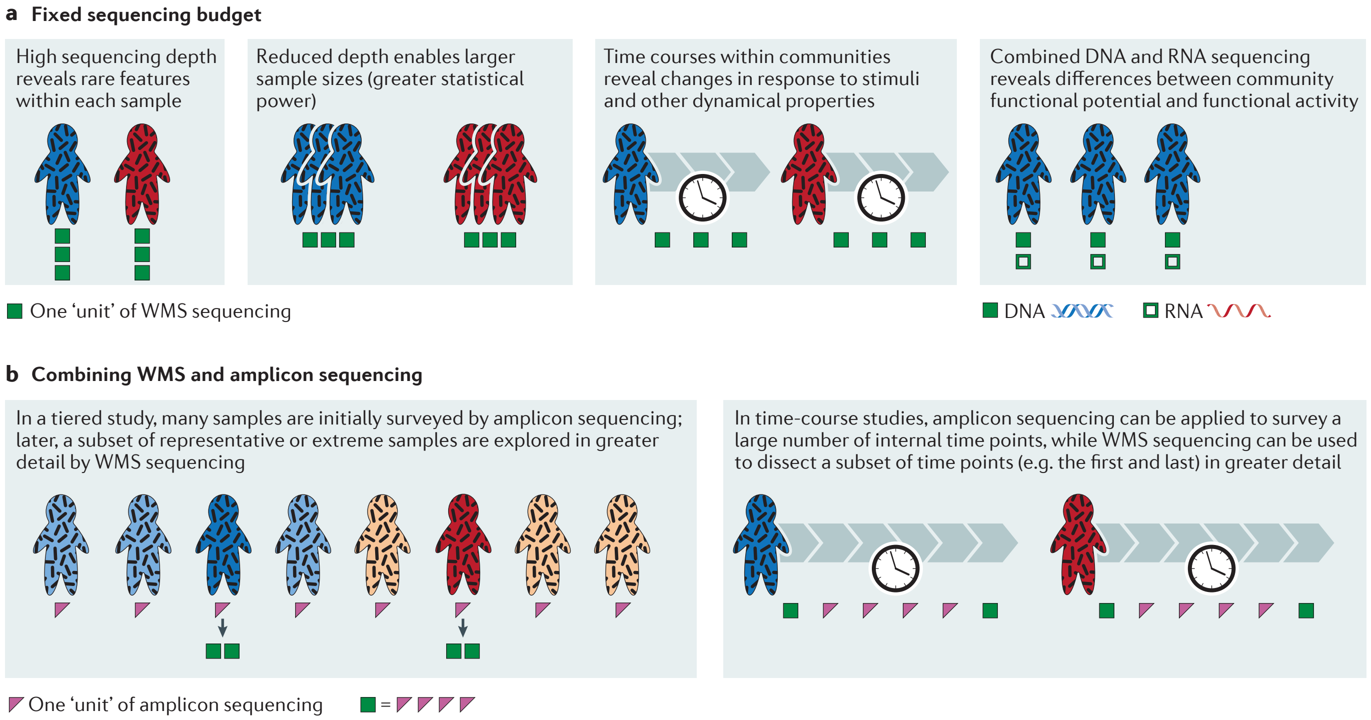
a Fixed sequencing budget
High sequencing depth reveals rare features within each sample
Reduced depth enables larger sample sizes (greater statistical power)
Time courses within communities reveal changes in response to stimuli and other dynamical properties
Combined DNA and RNA sequencing reveals differences between community functional potential and functional activity
🟩 One 'unit' of WMS sequencing
🟩 DNA
□ RNA
b Combining WMS and amplicon sequencing
In a tiered study, many samples are initially surveyed by amplicon sequencing; later, a subset of representative or extreme samples are explored in greater detail by WMS sequencing
In time-course studies, amplicon sequencing can be applied to survey a large number of internal time points, while WMS sequencing can be used to dissect a subset of time points (e.g. the first and last) in greater detail
◤ One 'unit' of amplicon sequencing
🟩 = ◤◤◤◤

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 that is achievable with a fixed sequencing budget (here, six 'units' of WMS sequencing). Greater sequencing depth facilitates the identification of rare species and rare variants of abundant species (such as single-nucleotide polymorphisms); 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 analyse 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 various high-throughput functional screens, including metaproteomics and metabolomics.

Low-error amplicon sequencing
(LEA–seq). An 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.

Reads
Short DNA or RNA sequences 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.

Taxonomic profiles based on standard amplicon sequencing are composed of operational taxonomic units (OTUs), which are often more specific than genera but in most cases are 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 used to study subspecies-level population structure in the vaginal microbiome[20] and to link sewage samples to specific faecal pollution sources[21]. In addition, a new, low-error approach to 16S rRNA gene sequencing, known as low-error amplicon sequencing (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 owing to its ability to identify variation throughout microbial genomes (FIG. 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, as a result of strain-specific gene-loss events, portions of these reference sequences may be absent in isolates of a species that is present in the sample, resulting in gaps in otherwise uniform coverage of the reference (FIG. 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[7]. Genomic islands were shown to contain multiple, functionally coherent genes (such as those encoding subunits

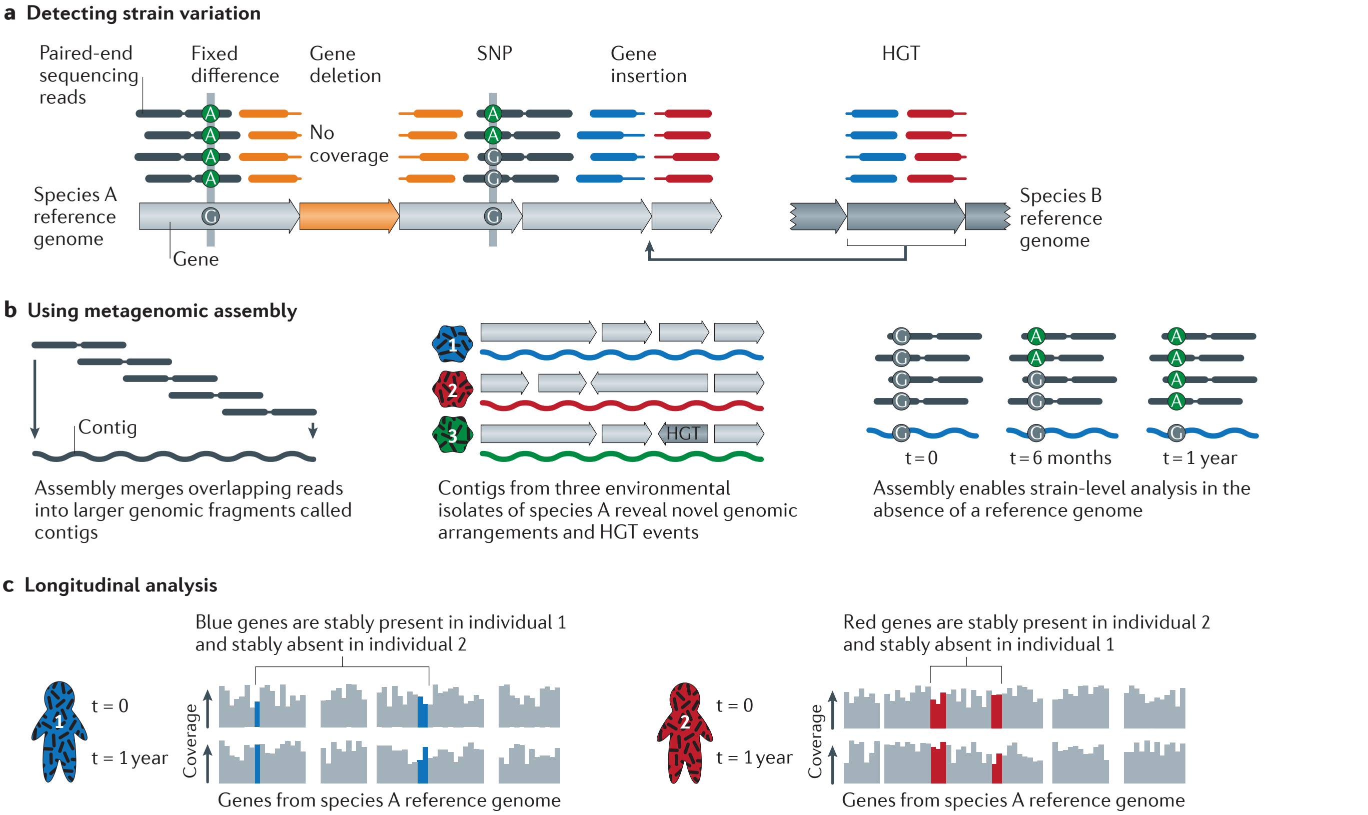
a Detecting strain variation
Paired-end sequencing reads
Fixed difference
Gene deletion
No coverage
SNP
Gene insertion
HGT
Species A reference genome
Gene
Species B reference genome
A
G
b Using metagenomic assembly
Contig
Assembly merges overlapping reads into larger genomic fragments called contigs
1
2
3
HGT
Contigs from three environmental isolates of species A reveal novel genomic arrangements and HGT events
t = 0
t = 6 months
t = 1 year
Assembly enables strain-level analysis in the absence of a reference genome
c Longitudinal analysis
Blue genes are stably present in individual 1 and stably absent in individual 2
t = 0
t = 1 year
Coverage
Genes from species A reference genome
Red genes are stably present in individual 2 and stably absent in individual 1
Genes from species A reference genome

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 4× coverage, represented by four 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 that is not found in the reference genome 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, analysing 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.

Single-nucleotide polymorphisms

(SNPs). Positions in a reference genome that occur in more than one nucleotide state (A, C, G and T) among the members of a population.

Contigs

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.

of the V-type H^+ ATPase) that were gained and lost together, suggesting a mechanism for individual-specific 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].

Missing genomic elements are detectable at relatively low WMS sequencing depths, but greater sequencing depths enable the confident detection of a wider variety of strain-level variants, including single-nucleotide polymorphisms (SNPs; FIG. 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 1 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 antibiotic-resistance genes were among the most variable, whereas housekeeping genes were among the most conserved.

In addition to enabling SNP identification, deeper WMS sequencing can 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 catalogues of microbial reference genomes (BOX 1). Indeed, it is increasingly possible to assemble whole microbial genomes from such communities and analyse 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; FIG. 2).

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 (that is, vertical transmission from a mother cell to daughter cell). HGT is also referred to as lateral gene transfer (LGT).

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 timescale of approximately 1 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 in which normal development is disrupted by medical intervention in early life[33]. For example, longitudinal WMS sequencing of an infant delivered by caesarean-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 carried out 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, showing 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 the human body (FIG. 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, although 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 (FIGS 1,2).

Multi-omic analyses

The preceding sections demonstrated that DNA sequence information can be used to profile microbial communities in several insightful but under-used ways. However, although the genomic content of a community describes its functional potential (what the community is capable of doing), it does not provide a direct measure of community functional activity (what the community is doing in a particular condition or at a particular time point). 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-omic data types such as transcriptomics, proteomics and metabolomics are needed.

Measuring functional activity with metatranscriptomics. Metatranscriptomics involves sequencing all of the RNA produced by a microbial community (the 'total RNA' of the community). It is crucial to enrich for microbial mRNAs by depleting rRNA before metatranscriptomic sequencing because mRNAs are vastly outnumbered by bacterial rRNAs in the total microbial RNA pool[40]. Microbial mRNA is then converted into 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 (FIG. 1).

Metatranscriptomic approaches were first applied to freshwater and marine microbial communities[41–43]. These studies demonstrated that, in a similar way to DNA, microbial total RNA could be used to profile community structure, function and diversity. Moreover, these studies showed that RNA sequencing produced

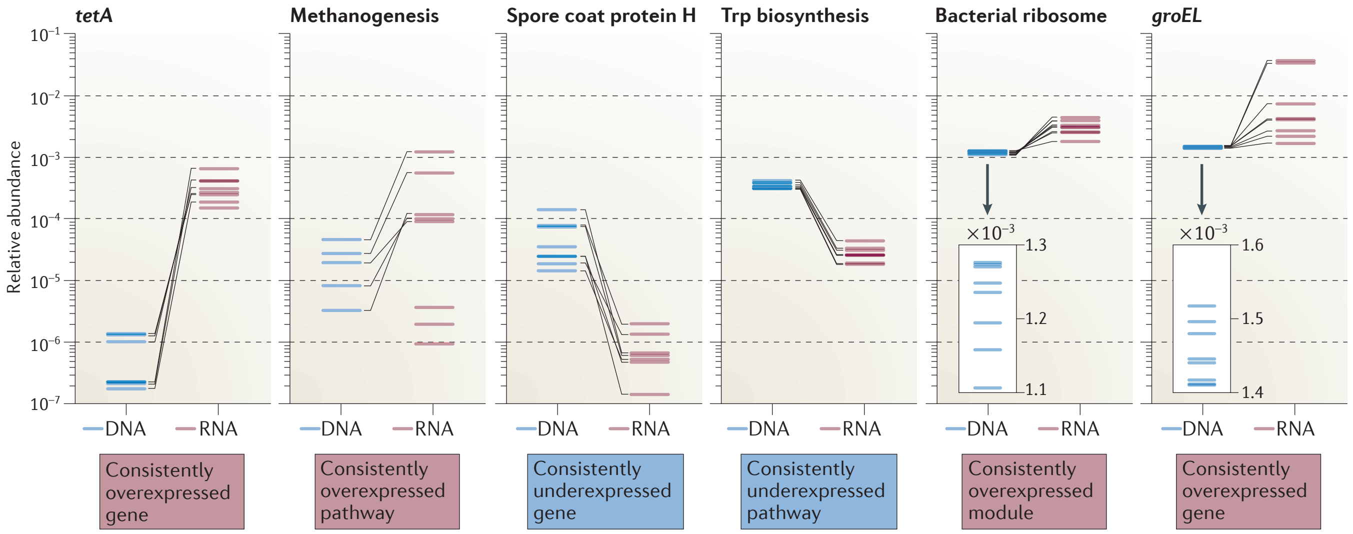

Figure 3 | **Relating the metatranscriptome and metagenome in the human gut.** In this example, shotgun RNA and DNA sequence data from gut microbiome samples of eight healthy individuals[25] were functionally profiled using the Human Microbiome Project Unified Metabolic Analysis Network (HUMAnN)[112]. Each panel shows 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), genes involved in methanogenesis (an important metabolic pathway among gut archaeal species) and the bacterial ribosome, and *groEL* (which encodes a bacterial chaperone protein) were strongly overexpressed because RNA related to these genes consistently exceeded the equivalent DNA in 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, bacterial ribosomal subunits were being continuously synthesized). Conversely, genes for spore coat protein H (which is involved in the response to nutrient starvation) and the synthesis of the amino acid tryptophan (Trp) were strongly underexpressed (DNA for these genes consistently exceeded the equivalent RNA in abundance). Reduced transcription of these genes suggests that these microbial functions were downregulated in the healthy human gut, which was likely to be a result of the high bioavailability of nutrients (including Trp) from the host's diet. Transcription of genes encoding bacterial ribosomal proteins and *groEL* was 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.

large amounts of novel sequence information, presumably by capturing organisms or genes of low copy number that are undersampled by DNA sequencing alone. In addition, metatranscriptomic sequencing provides a means to detect and quantify RNA viruses[44,45], which (except for integrated retroviruses[46]) are otherwise not included in DNA-based metagenomic surveys.

Combining metatranscriptomics with DNA-based taxonomic and functional profiling reveals prominent overexpression or underexpression of particular functions and, in some cases, the activities of whole organisms, 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 underexpressed (DNA is consistently found to be more abundant than the equivalent RNA) by bacteria in this environment, presumably because these compounds are readily available from the host and thus their synthesis by microorganisms would be energetically unfavourable[25] (FIG. 3). Sporulation was also strongly inactivated, presumably because bacteria are growing under ideal conditions in the healthy human gut (FIG. 3).

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

Outside 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 overexpression of putative virulence factors[47]. These findings suggest a role for metatranscriptomics in identifying bacterial

Sporulation
A stress response mechanism used by (primarily Gram-positive) bacteria to survive periods of nutrient depletion.

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 the degradation pathways targeting simpler, aliphatic hydrocarbons were highly expressed, whereas 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, in which 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 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 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 fermented milk product, particularly in pathways related to carbohydrate metabolism; such changes would go undetected in a metagenomics-only approach. Finally, whereas recent surveys of the human gut microbiome have revealed a remarkable degree of conservation in 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 (FIG. 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 in which the functional activity is regulated by 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, such as phosphorylation. Peptides are then associated with full-length proteins by sequence-homology-based searches against reference databases, in a similar way 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, more than 50% of total microbial proteins were involved in housekeeping functions, including translation and energy production[55]. Comparative metaproteomics of the gut microbiome in patients with Crohn 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 in patients with Crohn disease. Depletion of these compounds, which are proposed to play a part 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 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 the human body. For example, metaproteomics has been applied to monitor changes in biofilm formation in environmental communities that are 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 B12, 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

Microbiota
The collection of microorganisms (of all types: bacteria, archaea, viruses and eukaryotes) inhabiting a particular environment.

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, metaproteomic analysis has been applied to a waste-water-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 upregulation of ATPases); changes that first occurred late in the time course but were then maintained (termed long-term adaptation, which included the upregulation 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 with 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 high-performance liquid chromatography) to separate compounds, which are then identified and quantified using 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]), metabolomic methods are compatible with the sample biomass and preparations that are typical of meta-omic sequencing experiments.

Metabolomics shares several challenges with other meta-omic methods, including the fact that it analyses a broad catalogue of potential features that occur with high dynamic range, but it is further complicated by the non-uniformity of the features (molecules) that are profiled[64]. For example, whereas all transcripts belong to the same class of biomolecules (RNAs), metabolites span a broad range that includes small, hydrophilic carbohydrates (such as glucose), large, hydrophobic lipids (such as triacylglycerides) and 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 a good 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 antitumorigenic effects[65]. In the earlier example from Crohn disease, changes in SCFA profiles were inferred on the basis of 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-based 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 that led 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 blood 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 by-product of bacterial metabolism of cholesterol) were predicted to respond more favourably to statin therapy. In addition, recent findings indicate that the gut microbiota can directly metabolize xenobiotics ingested by their hosts into inactive forms, and that by-products 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–microorganism interaction and a second, dietary metabolite (arginine) that could mitigate the adverse effect.

One limitation of community metabolomic 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 (such as high salt concentrations) that make them less amenable to standard practices in metabolomics. However, a recent study proposed a metabolomics protocol that was specifically adapted for environmental samples through a stable isotope labelling step[71]. Samples from the environmental community of interest (in this case an acid mine drainage site) were cultured with ^{15}N-labelled ammonium sulphate 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 metagenomic analysis revealed that taurine was most likely to be 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.

Integrating multi-omic data

The studies and methods introduced in the preceding sections highlight many advantages of collecting additional multi-omic data types beyond DNA sequences in the characterization of microbial communities. RNA, proteins and metabolites all provide 'pictures' of the functional activity of a community, and these pictures often differ markedly from the functional potential that one would infer from DNA sequence alone (FIG. 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, when one must trade-off between the number of communities sampled and the number of multi-omic assays performed per sample (FIG. 1). One must also carefully consider the choice of analysis methods for integrating multi-omic data types, some of which will be general to all studies, whereas 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 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 (FIG. 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], the methanogenesis pathways of which were strongly overrepresented in the transcript pool relative to their metagenomic abundance (FIG. 3). Such insights would not have been possible if DNA or RNA data were considered 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 (FIG. 4). Such techniques have been widely applied to multi-omic data generated from model organisms 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].

These techniques have been developed for use in model organisms, but they are fully applicable to microbial communities, although they have not yet seen wide application. This is in some ways understandable owing to the complexity of microbial communities and the limited data that are available for analysis. Nevertheless, clear advantages of integrating independent multi-omic 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 disease and colorectal cancer) used two different multi-omic methods (proteomics and metabolomics) to show that disease conditions were associated with shifts in microbial metabolism of SFCAs[56,66]. Assuming that 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 under-used unless integrative techniques are used to combine data within and across studies.

Last, integrative techniques are crucial 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 (FIG. 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 because only a subset of *E. lenta* strains (those encoding the *cgr* operon) interacted in this manner. Metatranscriptomic data revealed a transcriptional

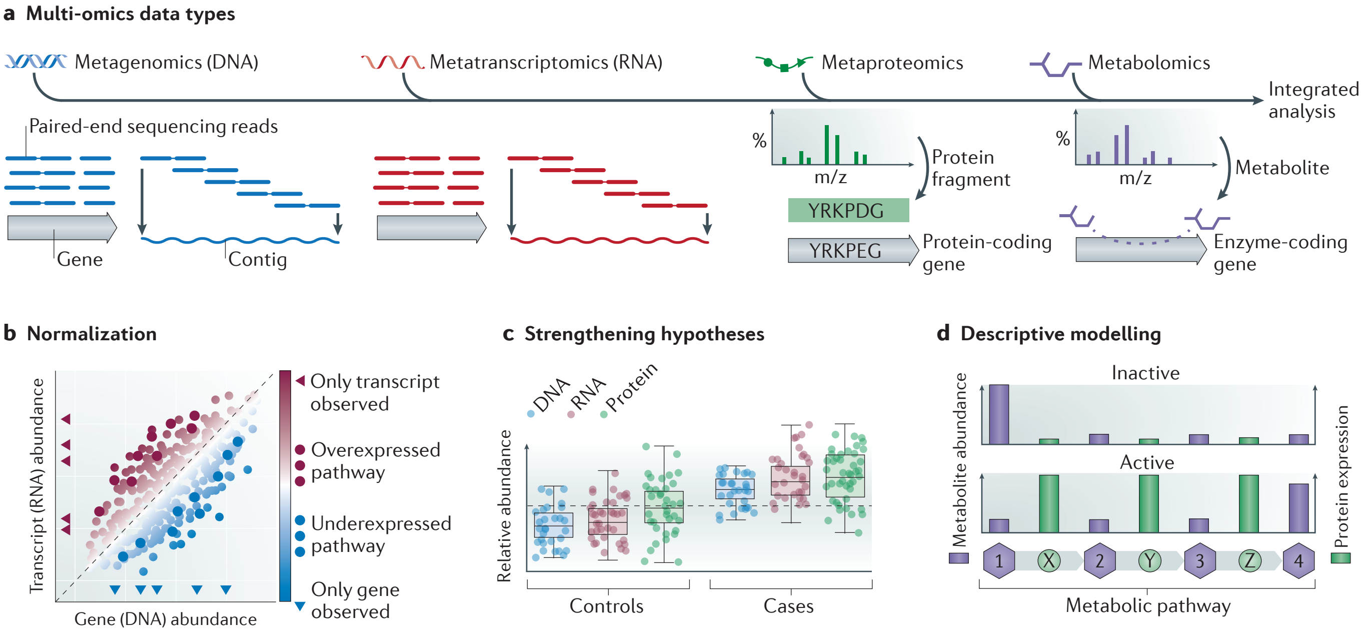

Figure 4 | **Integrating multi-omic data for deeper biological insights.** To facilitate integrated analysis of a microbiome sample, distinct multi-omic data types are often associated with microbial genes or gene families that act as a shared point of reference (part **a**). These genes may be taken 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 on the basis of 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 for and advantages in performing 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 modelling. For example, combining data from proteomic and metabolomic 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**). m/z, mass-to-charge ratio.

regulatory mechanism underlying the interaction: the *cgr* operon was upregulated in the presence of digoxin, but the upregulation was dampened in the presence of arginine. Achieving this level of detail in a descriptive model of a biological phenomenon depends crucially on the integration of multiple data types. Notably, identifying such mechanisms from high-dimensional multi-omic data requires special statistical techniques and considerations (BOX 2).

In addition to descriptive modelling, 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 REF. 78). Briefly, these methods involve the creation of a network in which metabolites are linked as reactants and products of enzymatic reactions. Constraint-based modelling (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 the level of enzyme expression[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 modelling 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 in which microorganisms produce crucial 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.

Flux balance analysis
A computational method for representing the steady-state metabolic network of an organism or community and evaluating its ability to produce a set of target metabolites from a set of input metabolites.

Box 2 | **Statistical considerations in multi-omic data integration**

Distinct multi-omic data types can be combined appropriately with exploratory, unsupervised approaches or by 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 (BOX 1). Common ordination methods include principal component analysis, principal coordinates analysis and non-metric multidimensional scaling (reviewed in an ecological context in REF. 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, principal coordinates analysis was used to generate a broad overview of hundreds of samples collected during the Human Microbiome Project[7]. In that case, it provided an efficient means of visualizing the ecological similarity of microbial communities from human skin and nasal samples relative to the more diverse communities found in oral, urogenital and gastrointestinal samples.

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 data sets to lower-dimensional spaces, as described above. The separate ordinations are then compared to see whether 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 are dependent on 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 (MWASs), which seek to link individual microbial features with other properties (such as disease status). MWAS analysis 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-normally distributed data and loss of power from performing many comparisons. Supervised methods that are appropriate for microbiome data (reviewed in REF. 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 LDA Effect Size (LEfSe)[118]). Supervised integration methods have been crucial for identifying metagenomic biomarkers in various human diseases, including obesity[8–10], inflammatory bowel disease[11–13] and cancer[14–16].

Conclusions drawn from such models will need to be further refined in the future by integrating additional multi-omic data types, such as using metabolomic profiling to validate proposed metabolic exchange relationships.

Summary and outlook

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 under-appreciated aspects of microbial community behaviour, such as strain-level phenomena, temporal dynamics and functional activity (FIGS 1,4). However, gaining a complete understanding of 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 the 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 also true of microbial isolates[86,87], raising the possibility that many gene products may actually be easier to characterize in communities because they might be functional in this context but inactive in laboratory monocultures. Likewise, few steps have yet been taken towards richer predictive models of microbial communities that incorporate regulatory relationships, ecology or inter-organismal signalling 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 before the development of accompanying mechanistic or ecological explanations. However, other diseases, such as inflammatory bowel disease[11–13], seem to be 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, fuelled 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.

1. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249 (1998).
2. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
3. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
One of the first large-scale environmental metagenomic sequencing projects; it presents profiles of taxonomic composition and function from geographically diverse marine microbial communities.
4. Rondon, M. R. *et al.* Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* **66**, 2541–2547 (2000).
5. Kembel, S. W. *et al.* Architectural design influences the diversity and structure of the built environment microbiome. *ISME J.* **6**, 1469–1479 (2012).
6. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
The first large-scale exploration of the human microbiome using metagenomic sequencing; it profiles 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 are shared across individuals.
7. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
The largest and most complete survey of the healthy human microbiome to date; it 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.
8. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
9. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
10. Ley, R. E. Obesity and the human microbiome. *Curr. Opin. Gastroenterol.* **26**, 5–11 (2010).
11. Manichanh, C. *et al.* Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut* **55**, 205–211 (2006).
12. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).
A metagenomic assessment of perturbations of the human gut microbiome in inflammatory bowel disease; it reveals that changes in functional composition are more pronounced than changes in community membership.
13. Berry, D. & Reinisch, W. Intestinal microbiota: a source of novel biomarkers in inflammatory bowel diseases? *Best Pract. Res. Clin. Gastroenterol.* **27**, 47–58 (2013).
14. Marchesi, J. R. *et al.* Towards the human colorectal cancer microbiome. *PLoS ONE* **6**, e20447 (2011).
15. Kostic, A. D. *et al.* Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
16. Tjalsma, H., Boleij, A., Marchesi, J. R. & Dutilh, B. E. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nature Rev. Microbiol.* **10**, 575–582 (2012).
17. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
An early metagenomic survey of the human gut microbiome that considers both stool and mucosal samples; it reveals many previously uncultured taxa along with strong inter-subject and inter-site differences.
18. Karch, H., Tarr, P. I. & Bielaszewska, M. Enterohaemorrhagic *Escherichia coli* in human medicine. *Int. J. Med. Microbiol.* **295**, 405–418 (2005).
19. Eren, A. M. *et al.* Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4**, 1111–1119 (2013)
This paper presents a computational approach for improving taxonomic resolution in surveys of microbial communities based on 16S rRNA sequencing.
20. Eren, A. M. *et al.* Exploring the diversity of *Gardnerella vaginalis* in the genitourinary tract microbiota of monogamous couples through subtle nucleotide variation. *PLoS ONE* **6**, e26732 (2011).
21. McLellan, S. L. *et al.* Sewage reflects the distribution of human faecal *Lachnospiraceae*. *Environ. Microbiol.* **15**, 2213–2227 (2013).
22. Faith, J. J. *et al.* The long-term stability of the human gut microbiota. *Science* **341**, 1237439 (2013).
This paper presents a sequencing method for improving taxonomic resolution in surveys of microbial communities based on 16S rRNA sequencing; the method was used 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. *Nature Methods* **9**, 811–814 (2012).
24. Scher, J. U. *et al.* Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife* **2**, e01202 (2013).
25. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).
26. Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
27. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–1665 (2012).
28. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotech.* **31**, 533–538 (2013).
29. Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Commun.* **4**, 2120 (2013).
30. Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife* **2**, e01102 (2013).
31. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048–1052 (2014).
32. Caporaso, J. G. *et al.* Moving pictures of the human microbiome. *Genome Biol.* **12**, R50 (2011).
33. Gronlund, M. M., Lehtonen, O. P., 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.* **28**, 19–25 (1999).
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.* **23**, 111–120 (2013).
35. Kong, H. H. *et al.* Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res.* **22**, 850–859 (2012).
36. Khoruts, A., Dicksved, J., Jansson, J. K. & Sadowsky, M. J. Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent *Clostridium difficile*-associated diarrhea. *J. Clin. Gastroenterol.* **44**, 354–360 (2010).
37. Turnbaugh, P. J. *et al.* The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Sci. Transl. Med.* **1**, 6ra14 (2009).
38. Spencer, M. D. *et al.* Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology* **140**, 976–986 (2011).
39. Rodriguez-Brito, B. *et al.* Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4**, 739–751 (2010).
40. Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, R23 (2012).
41. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).
42. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3**, e3042 (2008).
43. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA* **105**, 3805–3810 (2008).
44. Zhang, T. *et al.* RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* **4**, e3 (2006).
45. Culley, A. I., Lang, A. S. & Suttle, C. A. Metagenomic analysis of coastal RNA virus communities. *Science* **312**, 1795–1798 (2006).
46. Willner, D. *et al.* Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **4**, e7370 (2009).
47. Duran-Pinedo, A. E. *et al.* Community-wide transcriptome of the oral microbiome in subjects with and without periodontitis. *ISME J.* 1659–1672 (2014).
48. Mason, O. U. *et al.* Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* **6**, 1715–1727 (2012).
49. McNulty, N. P. *et al.* The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci. Transl. Med.* **3**, 106ra106 (2011).
50. Maurice, C. F., Haiser, H. J. & Turnbaugh, P. J. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell* **152**, 39–50 (2013).
51. Gilbert, J. A. *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* **5**, e15545 (2010).
52. Altelaar, A. F., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Rev. Genet.* **14**, 35–48 (2013).
53. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
54. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
55. Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **3**, 179–189 (2009).
56. Erickson, A. R. *et al.* Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS ONE* **7**, e49138 (2012).
57. Smith, P. M. *et al.* The microbial metabolites, short-chain fatty acids, regulate colonic T_{Reg} cell homeostasis. *Science* **341**, 569–573 (2013).
This study demonstrates that SCFAs, a common class of microbial metabolites, have an important role in co-adaptation between the gut microbiome and host immune system.
58. Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013).
59. Mosier, A. C. *et al.* Elevated temperature alters proteomic responses of individual organisms within a biofilm community. *ISME J.* **9**, 180–194 (2015).
60. Sowell, S. M. *et al.* Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* **3**, 93–105 (2009).
61. Morris, R. M. *et al.* Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* **4**, 673–685 (2010).
62. Lacerda, C. M., Choe, L. H. & Reardon, K. F. Metaproteomic analysis of a bacterial community response to cadmium exposure. *J. Proteome Res.* **6**, 1145–1152 (2007).
63. Turnbaugh, P. J. & Gordon, J. I. An invitation to the marriage of metagenomics and metabolomics. *Cell* **134**, 708–713 (2008).
64. Tang, J. Microbial metabolomics. *Curr. Genomics* **12**, 391–403 (2011).
65. Williams, E. A., Coxhead, J. M. & Mathers, J. C. Anti-cancer effects of butyrate: use of micro-array technology to investigate mechanisms. *Proc. Nutr. Soc.* **62**, 107–115 (2003).
66. Weir, T. L. *et al.* Stool microbiome and metabolome differences between colorectal cancer patients and healthy adults. *PLoS ONE* **8**, e70803 (2013).
67. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011).
Using a combination of integrated multi-omic analysis and experimental work in mice, these authors demonstrate a functional link between the metabolism of dietary compounds by the gut microbiome and the development of CVD.
68. Koeth, R. A. *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nature Med.* **19**, 576–585 (2013).

69. Kaddurah-Daouk, R. *et al.* Enteric microbiome metabolites correlate with response to simvastatin treatment. *PLoS ONE* **6**, e25482 (2011).

70. Haiser, H. J. *et al.* Predicting and manipulating cardiac drug inactivation by the human gut bacterium *Eggerthella lenta*. *Science* **341**, 295–298 (2013).
Through an integrated multi-omic analysis, these authors identify an operon in a member of the human gut microbiome community that is involved in degradation (and hence loss of efficacy) of the cardiac drug digoxin.

71. Mosier, A. C. *et al.* Metabolites associated with adaptation of microorganisms to an acidophilic, metal-rich environment identified by stable-isotope-enabled metabolomics. *mBio* **4**, e00484-12 (2013).

72. Karaoz, U. *et al.* Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc. Natl Acad. Sci. USA* **101**, 2888–2893 (2004).

73. Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B. & Botstein, D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA* **100**, 8348–8353 (2003).

74. Lee, I., Date, S. V., Adai, A. T. & Marcotte, E. M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).

75. Myers, C. L. *et al.* Discovery of biological networks from diverse functional genomic data. *Genome Biol.* **6**, R114 (2005).

76. Jansen, R. *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453 (2003).

77. Park, C. Y., Hess, D. C., Huttenhower, C. & Troyanskaya, O. G. Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *PLoS Comput. Biol.* **6**, e1001009 (2010).

78. Thiele, I. & Palsson, B. O. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protoc.* **5**, 93–121 (2010).

79. Durot, M., Bourguignon, P. Y. & Schachter, V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.* **33**, 164–190 (2009).

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.* **6**, e1000859 (2010).

81. Stolyar, S. *et al.* Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol.* **3**, 92 (2007).

82. Klitgord, N. & Segre, D. Environments that induce synthetic microbial ecosystems. *PLoS Comput. Biol.* **6**, e1001002 (2010).

83. Heinken, A., Sahoo, S., Fleming, R. M. & Thiele, I. Systems-level characterization of a host–microbe metabolic symbiosis in the mammalian gut. *Gut Microbes* **4**, 28–40 (2013).

84. Mazumdar, V., Amar, S. & Segre, D. Metabolic proximity in the order of colonization of a microbial community. *PLoS ONE* **8**, e77617 (2013).

85. Howe, A. C. *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl Acad. Sci. USA* **111**, 4904–4909 (2014).

86. Roberts, R. J. *et al.* COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Res.* **39**, D11–D14 (2011).

87. Harrington, E. D. *et al.* Quantitative assessment of protein function prediction from metagenomics shotgun sequences. *Proc. Natl Acad. Sci. USA* **104**, 13913–13918 (2007).

88. Gough, E., Shaikh, H. & Manges, A. R. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent *Clostridium difficile* infection. *Clin. Infect. Dis.* **53**, 994–1002 (2011).

89. Bhavsar, A. P., Guttman, J. A. & Finlay, B. B. Manipulation of host-cell pathways by bacterial pathogens. *Nature* **449**, 827–834 (2007).

90. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).

91. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* **19**, 1141–1152 (2009).

92. Segata, N. *et al.* Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666 (2013).
An in-depth review of computational methods in microbial community analysis.

93. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

94. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).

95. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12**, 1889–1898 (2010).

96. Knights, D., Parfrey, L. W., Zaneveld, J., Lozupone, C. & Knight, R. Human-associated microbial signatures: examining their predictive value. *Cell Host Microbe* **10**, 292–296 (2011).

97. McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4**, 63–72 (2007).

98. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods* **10**, 1196–1199 (2013).

99. Brady, A. & Salzberg, S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods* **8**, 367 (2011).

100. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

101. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).

102. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).

103. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**, D284–289 (2012).

104. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).

105. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

106. Langille, M. G. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotech.* **31**, 814–821 (2013).

107. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.* **42**, D459–D471 (2014).

108. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).

109. Markowitz, V. M. *et al.* IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.* **40**, D123–D129 (2012).

110. Meyer, F. *et al.* The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386 (2008).

111. Konwar, K. M., Hanson, N. W., Page, A. P. & Hallam, S. J. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14**, 202 (2013).

112. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).

113. Ramette, A. Multivariate analyses in microbial ecology. *FEMS Microbiol. Ecol.* **62**, 142–160 (2007).

114. Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc. Natl Acad. Sci. USA* **106**, 1374–1379 (2009).
An in-depth review of statistical procedures for identifying patterns in high-dimensional microbial community data.

115. McHardy, I. H. *et al.* Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome* **1**, 17 (2013).

116. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol. Rev.* **35**, 343–359 (2011).

117. White, J. R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* **5**, e1000352 (2009).

118. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

119. Tickle, T. L., Segata, N., Waldron, L., Weingart, U. & Huttenhower, C. Two-stage microbial community experimental design. *ISME J.* **7**, 2330–2339 (2013).

Acknowledgements

This work was funded in part by US National Institutes of Health grants R01HG005969 and U54DK102557 (to C.H. and R. J. Xavier); US National Science Foundation grant DBI-1053486 (to C.H.); US Army Research Office grant W911NF-11-1-0473 (to C.H.); and Danone Research grant PLF-5972-GD (to W. S. Garrett).

Competing interests statement

The authors declare no competing interests.