Shotgun metagenomics, from sampling to analysis

Christopher Quince[1,7], Alan W Walker[2,7], Jared T Simpson[3,4], Nicholas J Loman[5] & Nicola Segata[6]

Diverse microbial communities of bacteria, archaea, viruses and single-celled eukaryotes have crucial roles in the environment and in human health. However, microbes are frequently difficult to culture in the laboratory, which can confound cataloging of members and understanding of how communities function. High-throughput sequencing technologies and a suite of computational pipelines have been combined into shotgun metagenomics methods that have transformed microbiology. Still, computational approaches to overcome the challenges that affect both assembly-based and mapping-based metagenomic profiling, particularly of high-complexity samples or environments containing organisms with limited similarity to sequenced genomes, are needed. Understanding the functions and characterizing specific strains of these communities offers biotechnological promise in therapeutic discovery and innovative ways to synthesize products using microbial factories and can pinpoint the contributions of microorganisms to planetary, animal and human health.

High-throughput sequencing approaches enable genomic analyses ideally of all microbes in a sample, not just those that are amenable to cultivation. One such method, shotgun metagenomics, is the untargeted ('shotgun') sequencing of all ('meta-') microbial genomes 'genomics' present in a sample. Shotgun sequencing can be used to profile taxonomic composition and functional potential of microbial communities and to recover whole genome sequences. Approaches such as high-throughput 16S rRNA gene sequencing[1], which profile selected organisms or single marker genes, are sometimes referred to as metagenomics, but this is a misnomer, as they do not target the entire genomic content of a sample.

In the 15 years since it was first used, metagenomics has enabled large-scale investigations of complex microbiomes (C.Q., N.S. and N.J.L.)[2–7]. Discoveries enabled by this technology include the identification of environmental bacterial phyla with endosymbiotic behavior[8] and species that can carry out complete nitrification of ammonia[9,10]. Other notable findings include the widespread presence of antibiotic genes in commensal gut bacteria[11], tracking of human outbreak pathogens[4] (C.Q. and N.J.L.), the strong association of the viral[12] and bacterial[13] fractions of the microbiome with inflammatory bowel diseases, and the ability to monitor strain-level changes in the gut microbiota after perturbations such as those induced by fecal microbiome transplantation[14].

Here we discuss best practices for shotgun metagenomics studies, including identification and tackling of limitations, and provide an outlook for metagenomics in the future.

[1]Warwick Medical School, University of Warwick, Warwick, UK. [2]Microbiology Group, The Rowett Institute, University of Aberdeen, Aberdeen, UK. [3]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [5]Institute for Microbiology and Infection, University of Birmingham, Birmingham, UK. [6]Centre for Integrative Biology, University of Trento, Trento, Italy. [7]These authors contributed equally to this work. Correspondence should be addressed to N.S. (nicola.segata@unitn.it).

Received 17 August 2015; accepted 12 July 2017; published online 12 September 2017; corrected after print 12 September 2017; doi:10.1038/nbt.3935

A typical shotgun metagenomics study comprises five steps, after the initial study design: (i) the collection, processing and sequencing of the samples; (ii) preprocessing of the sequencing reads; (iii) sequence analysis to profile taxonomic, functional and genomic features of the microbiome; (iv) statistical and biological post-processing analysis, and (v) validation (**Fig. 1**). Numerous experimental and computational approaches are available to carry out each step, which means that researchers are faced with daunting choices. And, despite its apparent simplicity, shotgun metagenomics has limitations, owing to potential experimental biases and the complexity of computational analyses and their interpretations. We assess the choices and common problems that accompany each step.

Shotgun metagenomics study design

The steps involved in the design of hypothesis-based studies are outlined in **Supplementary Figure 1**, with specific recommendations summarized in **Supplementary Box 1**. Microbial content can vary across samples from the same environment, which complicates the detection of statistically significant and biologically meaningful differences among small sets of samples. It is therefore important to establish that studies are sufficiently powered to detect differences, especially if the effect size is small[15]. One useful strategy may be to generate pilot data to inform power calculations[16,17]. Alternatively, a two-tiered approach, in which shotgun metagenomics is carried out on a subset of samples that have been pre-screened with less expensive microbial surveys such as 16S rRNA gene sequencing, may be adopted (N.S.)[18].

Controls can be difficult to obtain, especially for samples from complex environments. This is particularly important for those studying the human microbiota, in which the resident microbial communities are influenced by multiple factors, such as host genotype[19], age, diet and environmental surroundings[20]. Where feasible, we recommend longitudinal studies that incorporate samples from the same habitat over time rather than simple cross-sectional studies that compare 'snapshots' of two sample sets[21]. Importantly, longitudinal studies do

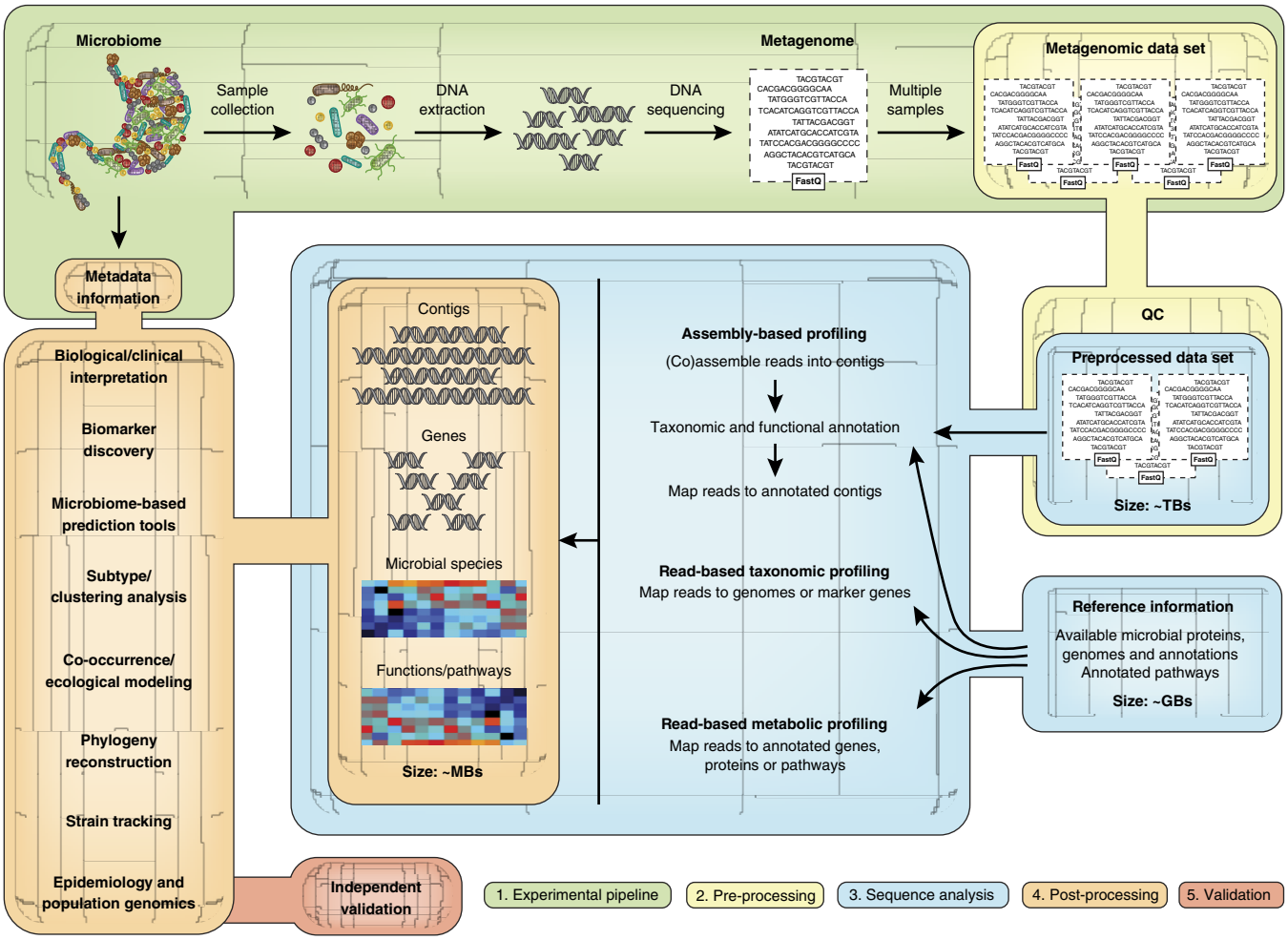

Figure 1 Summary of a metagenomics workflow. Step (1): study design and experimental protocol. The importance of this step is often underestimated in metagenomics. Step (2): computational pre-processing. Computational quality control (QC) steps minimize fundamental sequence biases or artifacts such as removal of sequencing adaptors, quality trimming, removal of sequencing duplicates (using for example, FastQC, Trimmomatic[121] or Picard tools). Foreign or non-target DNA sequences are also filtered, and samples are subsampled to normalize read numbers if the diversity of taxa or functions is compared. Step (3): sequence analysis. This should comprise a combination of 'read-based' and 'assembly-based' approaches depending on the experimental objectives. Both approaches have advantages and limitations (**Table 4**). Step (4): post-processing. Various multivariate statistical techniques can be used to interpret the data. Step (5): validation. Conclusions from high-dimensional biological data are susceptible to study-driven biases, so follow-up analyses are vital.

not rely on results from a single sample that might be a nonrepresentative outlier. Exclusion of samples that may be confounded by an unwanted variable is also prudent. For example, in studies of human subjects, exclusion criteria might include exposure to drugs that are known to affect the microbiome, such as antibiotics. If this is not feasible, then potential confounders should be factored into comparative analyses (**Supplementary Box 1**).

If samples originate in animal models, particularly co-housed rodents, the potential effects of animal age, housing environment[22,23] and even the sex of the person handling the animals[24] on microbial community profiles should be taken into account. It is usually possible to mitigate potential confounders in a study design by housing animals individually to prevent the spread of microbes between cage mates (although this may introduce behavioral changes, potentially resulting in different biases), housing animals derived from different experimental cohorts in the same cage or repeating experiments with mouse lines from different vendors or with different genetic backgrounds[25].

Finally, regardless of the type of sample being studied, it is crucial to collect detailed and accurate metadata. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) gene sequence (MIxS) standards were set out to provide guidance for required metadata[26], but metagenomics is now applied on such disparate types of environments that it is difficult to choose parameters that are suitable and feasible to obtain for every sample type. We recommend associating as much descriptive and detailed metadata as possible with each sample to make it more likely that comparisons between study cohorts or sample types can be correlated with a particular environmental variable[21].

Sample collection and DNA extraction

Sample collection and preservation protocols can affect both quality and accuracy of metagenomics data. Importantly, the effect size of these steps, in some circumstances, can be greater than the effect size of the biological variables of interest[27]. Indeed, variations in sample processing protocols can also be important confounders in

Table 1 Advantages and limitations of methods to enrich for microbial cells and DNA before sequencing

Technique	Advantages	Limitations
Whole-genome amplification[122]	• Generates sufficient DNA for sequencing from even tiny amounts of starting material • Can be applied directly to extracted environmental DNA • Can amplify DNA from the whole range of species present within a given sample	• Can introduce significant biases during amplification, which skew resulting metagenomics profiles • Chimeric molecules can be formed during amplification, which can confound the assembly step • Is unlikely to improve proportional abundance of DNA from a species of interest
Single-cell genomics[71]	• Can generate genomes from uncultured organisms • Can be combined with targeting approaches such as fluorescence *in situ* hybridization to select specific taxa, including those that might be rare members of the microbial community • Places genomic data within its correct phylogenetic context • Reference genomes can aid metagenomics assemblies	• Isolation of single cells can be expensive, requiring specialist equipment • Requires a whole-genome amplification step • Introduces biases during genome amplification so that it is usually possible to recover only partial genomes • Is prone to contamination
Flow sorting[123]	• Offers high-throughput means to sort cells of interest • Can select specific taxa, including those that might be rare members of the microbial community	• Requires expensive equipment and specialist operators • Requires intact cells • May not recover cells in the sample that are attached to surfaces or fixed in structures, such as biofilms • Is limited by flow rates and sort volumes in the number of cells that can be collected
In situ enrichment[124]	• Simplifies microbial community structure and can make it easier to assemble genomes from metagenomics data. • Presence of particular taxa within enriched samples can give clues about their functional roles within the microbial community.	• Requires that cells of interest can be maintained stably in a microcosm over the entire enrichment period. • Simplifies microbial community structure; biases results in favor of organisms that were able to thrive within the microcosm.
Culture or microculture[70]	• Allows extensive testing of isolates for phenotypic features • Reference genomes can aid metagenomics assemblies • Provides functional data to improve metagenomics annotations • Places genomic data within its correct phylogenetic context	• Is low throughput, can be highly labor intensive. • Is limited by difficulty of culturing some microbes in the laboratory • Is unlikely to recover rarer members of a microbial community, as cultured isolate collections will be dominated by the most abundant organisms
Sequence capture technologies[125]	• Can use oligonucleotide probes to identify species of interest, as recently demonstrated for culture-independent viral diagnostics • Can achieve higher sensitivity by focusing only on species of interest, particularly when large amounts of host contamination are present	• Uses kits that can be expensive • Like PCR, fails when target organisms vary compared to the reference sequences used to design the probes • Can give uneven genome coverage of targeted organisms, affecting assemblies
Immunomagnetic separation[126]	• Can enrich specific taxa, including those that might be comparatively rare members of the microbial community • Is far less expensive than techniques such as single-cell genomics or flow sorting • Is less technically challenging and time consuming than other targeted enrichment techniques	• Requires intact cells • Requires a specific antibody for the target cells of interest • May require whole-genome amplification after cell separation if target cell numbers are low
Background (human or eukaryotic) depletion techniques[127]	• Is particularly useful for samples where microbial cell numbers are much lower than eukaryotic cells • Has enhanced detection of microbial genomic data • Requires lower sequence depth to obtain good coverage of microbial genomes, reduced sequencing costs • Relatively inexpensive, not technically challenging	• Can lose bacterial DNA of interest during processing steps and bias subsequent microbiome profiling • May introduce contamination

meta-analyses of data sets from different studies (**Supplementary Box 1**). Collection and storage methods that have been validated for one sample type cannot be assumed to be optimal for other sample types. As such, careful preliminary work to optimize processing conditions for sample types is often necessary (**Supplementary Fig. 1**).

Key objectives are to collect sufficient microbial biomass for sequencing and to minimize contamination of samples. Enrichment methods can be used for environments in which microbes are scarce (**Table 1**). However, these procedures can introduce bias into sequencing data[28]. Several studies have shown that factors such as length of time between sample collection and freezing (A.W.W.)[29] and the number of freeze–thaw cycles a sample undergoes can affect the microbial community profiles that are detected; as such, collection and storage protocols and conditions should be recorded (**Supplementary Box 1**).

DNA extraction methodology can affect the composition of downstream sequence data[30]. The extraction method must be effective for diverse microbial taxa; otherwise, sequencing results may be dominated by DNA derived only from easy-to-lyse microbes. DNA extraction methods that include mechanical lysis (or bead beating) are often considered superior to those that rely on chemical lysis[31]. However, approaches based on bead beating vary in efficiency (A.W.W.)[32]. Vigorous extraction techniques, such as bead beating, can result in shortened DNA fragments, which can contribute to DNA loss during library preparation methods that use fragment size–selection techniques.

Contamination can occur during sample processing stages. Kit or laboratory reagents may contain variable amounts of microbial contaminants[33]. Metagenomics data sets from low-biomass samples (for example, skin swabs) are particularly vulnerable to this problem, because there is less 'real' signal to compete with low levels of contamination (A.W.W. and N.J.L.)[34]. We advise researchers working with low-biomass samples to use ultraclean reagents[35] and to incorporate 'blank' sequencing controls, in which reagents are sequenced without adding sample template (A.W.W. and N.J.L.)[34]. Other sources of contamination are cross-over from previous sequencing runs, presence of the PhiX control DNA that is typically used in Illumina-based sequencing protocols, and human or host DNA.

Table 2 Comparative evaluation of metagenomic assembly on mock microbial communities with known composition

Data set	Metagenomic assembly method	Assembly statistics for contigs >1 kb				
		Number of contigs ($\times 10^3$)	Total assembly size (Mb)	Reconstruction percentage	N50 ($\times 10^3$)	Percentage identity
Env. mock community[54]	MetaSPAdes	16.22 (11.26)	150.47 (108.39)	80.93 (58.30)	26.46 (25.88)	99.86 (99.96)
	MEGAHIT	21.82 (16.67)	146.72 (124.67)	78.91 (67.05)	16.94 (17.94)	99.93 (99.98)
HMP mock community[2]	MetaSPAdes	0.72 (0.42)	62.67 (31.95)	95.15 (48.50)	260.45 (178.28)	99.98 (99.99)
	MEGAHIT	1.43 (1.14)	62.09 (54.56)	94.27 (82.84)	124.02 (113.11)	99.99 (99.99)

Assemblies were produced using SPAdes (version 3.7.1) and MegaHIT (1.0.4) using the default recommended parameters for metagenomic assembly ("--meta" "-k 21,33,55,77" for SPAdes and "--presets meta-sensitive" for MegaHIT). The input metagenomes are Illumina sequencing of previously described mock communities[54] that were subsampled to 50 million reads for comparability. The subsampled paired-end fastq files are available at https://mgexamples.s3.climb.ac.uk/HMP_MOCK_SRR2726667_8.25M.1.fastq.gz and https://mgexamples.s3.climb.ac.uk/HMP_MOCK_SRR2726667_8.25M.2.fastq.gz for the HMP mock community (comprising 20 strains) and at https://mgexamples.s3.climb.ac.uk/MOCK_M63H.25M.1.fastq.gz and https://mgexamples.s3.climb.ac.uk/MOCK_M63H.25M.2.fastq.gz for the environmental (env.) mock community (comprising 59 strains). Reconstruction percentage and percentage identity were computed by mapping with BLASTN[128] the contigs against the genomes of the organisms in the mock communities. The N50 value corresponds to the size of the contig for which longer contigs represent at least half of the total assembly, and it is one of the key parameters for assessing the quality of an assembly. In parenthesis we report the statistics referred to 'perfect contigs' which are those contigs reconstructed by metagenomic assembly that have a match with >99% identity with the reference genome over the full length of the contig. Notably, 'perfect contigs' excludes chimeric contigs.

Library preparation and sequencing

The choice of library preparation and sequencing method rests on availability of materials and services, cost, ease of automation and DNA sample quantification. The Illumina platform is predominant in shotgun metagenomics owing to its wide availability, very high outputs (up to 1.5 Tb per run) and high accuracy (with a typical error rate of 0.1–1%), although the competing Ion Torrent S5 or S5 XL instruments are an alternative. Long-read sequencing technologies such as the Oxford Nanopore MinION and Pacific Biosciences Sequel have scaled up output, now reliably generating up to 10 Gb per run, so these platforms may therefore soon start to see adoption for metagenomics studies.

Given the very high outputs achievable in a single run, multiple metagenomic samples are usually sequenced at once by multiplexing of up to 96 or 384 samples, typically using dual indexing barcode sets available for all library preparation protocols. The Illumina platforms have problems with carryover (between runs) and carry-between (within runs)[36]. Recently, concern has been raised that newer Illumina instruments using a new amplification method (ExAmp) suffer from high rates of 'index hopping', where incorrect barcode identifiers are incorporated into growing clusters[37], but the extent of this problem on typical metagenomics projects has not been evaluated, and Illumina has suggested best practices to mitigate it. Researchers can evaluate the extent of such issues through randomly chosen control wells containing known spiked-in organisms as positive controls and template negative controls. Such measures are particularly critical for diagnostic metagenomics projects, where small numbers of pathogen reads may be a signal of infection against a background of high host contamination. Although still uncommon in the field, technical replicates would be useful to assess variability, and subjecting even a subset of samples to replication may give enough information to disentangle technical from true variability.

Multiple methods are available for the generation of Illumina sequencing libraries. These are usually distinguished by the method of fragmentation used. Transposase-based 'tagmentation', used in the Illumina Nextera and Nextera XT products, for example, is popular owing to its low cost ($25–40 per sample, and dilution methods can potentially reduce these costs further)[38]. Tagmentation approaches require small DNA inputs (1 ng DNA is recommended, but smaller amounts can be used). Such low inputs are possible owing to a subsequent PCR amplification step. However, as tagmentation targets specific sequence motifs, it may introduce amplification biases along with the well-known GC content biases associated with PCR. One way to reduce these biases is to use a PCR-free method relying on physical fragmentation (e.g., PCR-free TruSeq) to produce a sequencing library that may be more representative of the underlying species composition in a sample[39].

There are no published guidelines for the 'correct' amount of coverage for a given environment or study type, and it is unlikely that such a figure exists. As a rule of thumb, we often recommend choosing a system that maximizes output in order to retrieve sequences from as many low-abundance members of the microbiome as possible. Illumina HiSeq 2500 or 4000, NextSeq and NovaSeq produce high volumes of sequence data (between 120 Gb and 1.5 Tb per run) and are well suited for metagenomics studies (with the caveat about index hopping). The throughput per run of these instruments is known and, by deciding the level of multiplexing, investigators can set the desired per-sample sequencing depth. Typical experiments in 2017 aim to generate between 1 and 10 Gb, but these depths may be either excessive or insufficient, depending on the sensitivity required to detect rare members of a sample.

The Illumina platforms differ mainly in their total output and maximum read length. The Illumina HiSeq 2500, although now two generations old, is a popular choice for shotgun metagenomics, as it is able to generate 2 × 250-nt reads in rapid-run mode (generating up to 180 Gb per flow cell) or up to 1 Tb in high-output mode, with 2 × 125-nt reads. The newer HiSeq 3000 and 4000 systems further increase the overall throughput of a run (up to 1.5 Tb for the 4000) but are limited to read lengths of 150 nt. The NextSeq bench top instrument has similar output to that of the HiSeq 2500 rapid-run mode but is limited to reads of 150 nt. However, the NextSeq costs less than half the price of the HiSeq so may be attractive to research groups wishing to operate their own instruments. The recently released NovaSeq platform promises up to 3 Tb per flow cell in the near future. The Illumina MiSeq is limited by output (up to 15 Gb in 2 × 300 mode) but remains the *de facto* standard for single-marker-gene microbiome studies. The MiSeq (or MiniSeq) may still be useful for sequencing a limited number of samples or to assess library concentrations and barcode pool balancing, providing confidence of good results before running on the higher-throughput instruments, where individual runs may cost >$10,000.

Metagenome assembly

Numerous approaches for computationally reconstructing microbial community composition from a pool of sequence reads have been published. Choosing the 'best' is a daunting task and depends largely on the aims of the study.

Metagenome *de novo* assembly is conceptually similar to whole-genome assembly (J.S.)[40]. The de Bruijn graph approach[41] is currently a very popular metagenome assembly method. For single draft genome assemblies, a de Bruijn graph is constructed by breaking each sequencing read into overlapping subsequences of a fixed length *k*. This set of overlapping '*k*-mers' defines the vertices and edges of the

Table 3 Comparative evaluation of metagenomic assembly of a set of metagenomes from diverse environments

Sample	Assembler	Number of genes	Number of matches against nr (95% identity)	Number of species observed (nr at 95% identity)	Median number of single core genes	Number of annotated COGs	Number of annotated KEGG orthologs
Env. mock community[54]	MetaSPAdes	164,750	154,403	103	49.5	100,681	91,376
	MEGAHIT	164,146	154,185	105	49	97,119	91,035
HMP mock community[2]	MetaSPAdes	62,850	61,362	30	20	44,625	36,082
	MEGAHIT	63,304	61,617	38	20	44,289	36,394
Gut sample[2]	MetaSPAdes	169,399	111,119	365	44.5	79,414	76,500
	MEGAHIT	166,289	109,777	381	41.5	77,666	75,020
Ocean sample[6]	MetaSPAdes	124,251	7,397	118	42	51,138	68,633
	MEGAHIT	151,627	7,987	110	60.5	67,979	87,344
Soil sample[129]	MetaSPAdes	34,118	7,411	86	4	10,448	15,312
	MEGAHIT	44,396	11,008	132	11.5	17,671	22,524

Assemblies were produced using SPAdes and MegaHIT as reported in **Table 2**. The gut sample was sequenced by the HMP[2] (50 million reads subsampled metagenome at https://mgexamples.s3.climb.ac.uk/HMP_GUT_SRS052697.25M.1.fastq.gz and https://mgexamples.s3.climb.ac.uk/HMP_GUT_SRS052697.25M.2.fastq.gz), the soil sample by Ofek-Lalzar et al.[129] (50 million reads subsampled metagenome at https://mgexamples.s3.climb.ac.uk/SOIL_NATCOMM.25M.1.fastq.gz and https://mgexamples.s3.climb.ac.uk/SOIL_NATCOMM.25M.2.fastq.gz), and the ocean sample by Sunagawa *et al.*[6] (50 million reads subsampled metagenome at https://mgexamples.s3.climb.ac.uk/TARA_OCEAN.25M.1.fastq.gz and https://mgexamples.s3.climb.ac.uk/TARA_OCEAN.25M.2.fastq.gz). Functional annotations were performed as previously described[60] (C.Q. and N.J.L.), with the total number of genes identified from the assembled contigs using Prodigal[130] (run with default parameters and the "-p meta" flag), matches against the NCBI nonredundant (nr) data set using DIAMOND[131] at 95% identity, the single core genes from the 36 universal COGs[60], COGs[132] annotated using rpsblast[133] at *e*-value < 0.00001 and KEGG orthologs[94] with DIAMOND using blastp and default parameters.

de Bruijn graph. The assembler's task is to find a path through the graph that reconstructs the genome(s). This task is complicated by sequencing errors, which generate nongenomic sequences, and repetitive sequence, which can cause misassemblies and fragmentation of the assembly.

Metagenome assembly presents unique challenges. First, when assembling a single genome, it is typically assumed that sequence coverage along the genome will be approximately uniform. An assembler can use sequence coverage to identify repeat copies, distinguish true sequence from sequencing errors (J.S.)[42] and identify allelic variation[43]. Metagenome assembly is more difficult, because the coverage of each constituent genome depends on the abundance of each genome in the community. Low-abundance genomes may end up fragmented if overall sequencing depth is insufficient to form connections in the graph. Using a short *k*-mer size in graph formation can assist in recovering lower-abundance genomes, but this comes at the expense of increased frequency of repetitive *k*-mers in the graph, obscuring the correct reconstruction of the genomes. The assembler must strike a balance between recovering low-abundance genomes and obtaining long, accurate contigs for high-abundance genomes. A second problem is that a sample can contain different strains of the same bacterial species. These closely related genomes can cause branches in the assembly graph where they may differ by a single nucleotide variant or by the presence or absence of an entire gene or operon. The assembler will often stop at these branch points, resulting in fragmented reconstructions.

Metagenome-specific assemblers try to overcome these challenges. Meta-IDBA[44] uses a multiple *k*-mer approach to avoid the task of choosing a *k*-mer length that works well for both low- and high-abundance species. Meta-IDBA has extensions to partition the de Bruijn graph (as does MetaVelvet[45]), and the latest version, IDBA-UD, optimizes the reconstruction for uneven sequence-depth distributions[46]. The SPAdes assembler[47] has been extended for metagenome assembly and can be used for assembling libraries sequenced with different technologies (hybrid assembly).

For complex samples that are likely to contain hundreds of strains, the sequencing depth must be increased as much as possible. Computational time and memory may be insufficient to complete such assemblies. Distributed assemblers (J.S.)[48] such as Ray, which spread memory load over a cluster of computers, have been used to assemble metagenomes from human fecal samples[49]. To help assemble very complex samples, Pell *et al.*[50] developed a lightweight method to partition a metagenome assembly graph into connected components that can be assembled independently. Another method, latent strain analysis, partitions reads using *k*-mer abundance patterns, which enables assemblies of individual low-abundance genomes using a limited amount of memory[51]. MEGAHIT uses succinct data structures to reduce the memory requirements of assembling complex metagenomes and achieves very quick run times[52].

There is little community consensus on how well different assemblers perform with respect to key metrics such as completeness, continuity and propensity to generate chimeric contigs. Despite metagenomic analysis 'bake-offs' aimed at making concrete recommendations for analysis software, it is likely that software performance depends on biological factors (e.g., underlying microbial community structure) and technical factors (e.g., sequencing platform characteristics and coverage). This effect was observed at an Assemblathon (J.T.S.)[53], where no single assembler came out 'best'.

We analyzed assembly results from mock synthetic and real communities (**Tables 2** and **3**). We evaluated MEGAHIT[52] and metaSPAdes[47] for their abilities to reconstruct known genomes from the mock communities and capture taxonomic and gene diversity in the real data sets. Both successfully reconstructed more than 75% of the mock communities—one comprising 20 organisms[2], the other, 49 bacterial and 10 archaeal species (C.Q.)[54]. MetaSPAdes generated longer contigs, but these appeared to be less accurate. When restricted to contigs that exactly matched the references in the mock community, MEGAHIT succeeded in reconstructing more of the true genomes. Choice of assembler in this case would therefore depend on the relative importance of contig size versus accuracy. Across the true data sets (**Table 3**), consistent patterns were hard to discern. However, examination of median single-copy core gene number (to estimate the number of genomes in the assembly) suggests that for the more complex soil and ocean communities, MEGAHIT assembled more genes that could be functionally annotated. The key lesson here is that different state-of-the-art programs will be optimal for different data sets while requiring similar run times (about 48 h using 16 threads on the largest sample) and main memory usage (not exceeding 125 GB). It is prudent, therefore, to attempt more than one assembly approach. The CAMI challenge reported that MEGAHIT was in

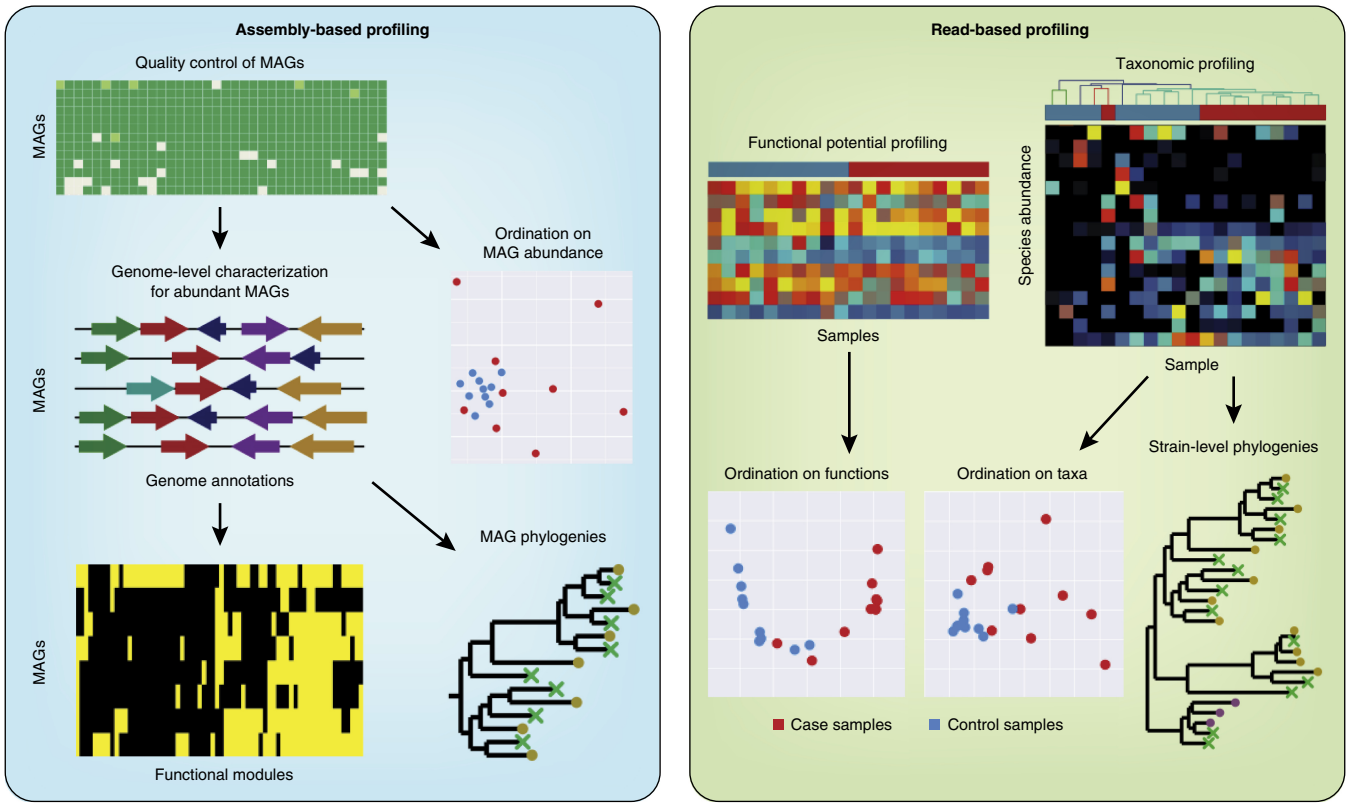

Figure 2 Assembly-based and assembly-free metagenome profiling. Starting from a metagenomic case-control design, we illustrate some of the steps needed to identify the organisms and the encoded functions and to try to links these samples' characteristics with the case or control condition. Left, an assembly-based pipeline (this can be fully reproduced following the commands and the code provided as a GitHub repository at https://github.com/chrisquince/metag-rev-sup). Right, a read-based pipeline using MetaPhlAn2 (ref. 87), HUMAnN2 (ref. 93) and a recent strain-level extension of the MetaPhlAn2 approach[87]. (The raw data are available at http://metagexample.s3.climb.ac.uk/Reads.tar.gz.)

the top three metagenomics assemblers across their benchmark data sets (C.Q.)[55] and, together with metaSPAdes (not evaluated in CAMI), is probably the best current choice. Whatever assembler is used, the result will not be genomes but rather potentially millions of contigs, and this motivates the need for binners to link the contigs back to the genomes they derived from.

Binning contigs

Metagenome assemblies are highly fragmented, comprising thousands of contigs (**Table 2**), and researchers do not know *a priori* which contig derives from which genome, or even know how many genomes are present. The aim of contig 'binning' is to group contigs into species. Supervised binning methods use databases of already sequenced genomes to label contigs into taxonomic classes. Unsupervised (clustering) methods look for natural groups in the data.

Both supervised and unsupervised methods have two main elements: a metric to define the similarity between a given contig and a bin, and an algorithm to convert those similarities into assignments. For taxonomic classification, contig homology against known genomes is a potentially useful approach, but most microbial species have not been sequenced, so a large fraction of reconstructed genomic fragments cannot be mapped to reference genomes. This has motivated the use of contig sequence composition for binning. Different microbial species' genomes contain particular combinations of bases, and this results in different *k*-mer frequencies[56]. Metrics based on these *k*-mer frequencies can be used to bin contigs, with tetramers considered the most informative for binning of metagenomics data[57]. Many software choices based on these frequencies are available, such as naive Bayes classifiers[58] or support vector machines[59], but sequence composition often lacks the specificity necessary to resolve complex data sets to the species level in complex communities (C.Q. and N.J.L.)[57,60,61].

Clustering of contigs is appealing because it does not require reference genomes. Until recently, most contig clustering algorithms, such as MetaWatt[61] and SCIMM[62], used various species composition metrics, sometimes coupled with total coverage. Recently, as multisample metagenome data sets have been produced, researchers have realized that contig coverage across multiple samples provides a much more powerful signal to group contigs together[63,64]. The underlying principle is that contigs from the same genome will have similar coverage values within each metagenome, although intra-genome GC content variation and increased read depth around bacterial origins of replication can challenge this assumption[65]. For example, the first algorithms, such as extended self-organizing maps[63], required human input to perform clustering, which is based on coverage information and composition that could be visualized in two dimensions[64]. Completely automated approaches, such as CONCOCT (C.Q. and N.J.L.)[60], GroopM[66] and MetaBAT[67], are now available, and they are convenient for large data sets, but better results may be obtained when combined with human refinement, for instance using a visualization tool such as Anvio (C.Q.)[68].

Methods for reconstructing metagenomic assembled genomes (MAGs) are indispensable to uncover the diversity of bacteria. The recovery of nearly 1,000 MAGs from candidate phyla, with no cultured representatives, from acetate-enriched and filtered groundwater samples showcased the potential of this approach[8]. Recovered genomes were small, with minimal metabolism, and formed a monophyletic clade that is separate from the previously cultured diversity of bacteria. These have been proposed as a new bacterial subdivision, the Candidate Phyla Radiation, revealed through metagenomics[69].

Completeness of MAGs is usually evaluated by examination of single-copy core genes, found in most microbial genomes, such as tRNA synthetases or ribosomal proteins. A pure MAG will have all these genes present in single copies. Once constructed, the MAGs provide a rich data set for comparative genomics, including the construction of phylogenetic

Box 1 Limitations and opportunities in metagenomics

There are several limitations and challenges to shotgun metagenomic studies. Limitations include:

'Entry-level access'. It is still expensive to sequence and analyze large numbers of metagenomes without access to sequencing and computational facilities. Improved sequencing platforms and cloud computing facilities should decrease these entry-level costs.

Comprehensiveness of genome catalogs. The set of >50,000 microbial genomes available is biased toward model organisms, pathogens and easily cultivable bacteria. All metagenomic computational tools rely to some extent on available genomes, and they are thus affected by the biases in the reference sequence resources.

Biases in functional profiling. Profiling of the functional classes present in a metagenome is hindered by the lack of validated annotations for most genes, an issue that can be mitigated only by expensive and low-throughput gene-specific functional studies. Moreover, intrinsic microbiome properties, such as its average genome size, can critically affect quantitative profiling[134].

Microbial dark matter. Several members of a microbiome might have not been characterized before with culture-based methods or with metagenomics. Assembly-based approaches can recover part of this 'microbial dark matter'. A fraction of reads may still remain unused after assembly, and the size of this fraction is highly dependent on community structure and complexity (**Tables 2** and **3**). It is also affected by features such as sequencing noise, contaminant DNA and microbes and plasmids that remain taxonomically obscure even after parts of their genomes are assembled.

'Live or dead' dilemma. DNA persists in the environment after the death of the host cell, so sequencing results may not be representative of the active microbial population. Compounds such as propidium monazide, which binds free DNA, as well as DNA within dead or damaged cells, or techniques such as metatranscriptomics, may be used if the aim is to study the active microbes.

'Curse of compositionality'. Quantitative metagenomic features are reported as fractional values without links to the real absolute concentration. Variations in the true concentration of organisms across samples can thus produce false correlations. For example, if a highly abundant organism doubles its concentration in two otherwise identical samples, all the other organisms in the sample will appear to be differentially abundant after normalization.

Mucosa-associated microbiome sequencing. Human mucosal tissues are crucial interfaces between microbes and the immune system, but sequencing the mucosal microbiome with shotgun metagenomics is very challenging due the extremely high fraction of human DNA and the low microbial biomass.

Shotgun metagenomic studies also present opportunities, such as:

Integrative meta-omics. Although complementing DNA sequencing with RNA, protein and metabolomic high-throughput assays is possible with shotgun metatranscriptomics, mass-spectrometry-based metaproteomics and metabolomics[73], it is unclear how to integrate and analyze meta-omic data within a common framework.

Virome shotgun sequencing. Viral organisms can be detected by shotgun metagenomics, but virome enrichment techniques are usually needed to access a broader set of viruses. Virome analysis is also computationally challenging because of limited availability of viral genomes and a lack of inter-family phylogenetic signals.

Strain-level profiling. The genomic resolution of single isolate sequencing is still higher than what can be achieved for single organisms in a metagenomic context. Increasing the profiling resolution to the level of single strains would be crucial for in-depth population genomics and microbial epidemiology.

Longitudinal study design. Many shotgun metagenomic studies are cross-sectional and thus unpowered for assessing inter- versus intra-subject variability and microbiome temporal variation. Tools for longitudinal settings have been developed[60], but more methods and data are needed to investigate the temporal dimension[135].

Disentangling cause from effect. Hypotheses from metagenomic studies should be followed up with experimental work to validate correlations and associations. Longitudinal and prospective settings can potentially provide direct insights into the causative dynamics of conditions of interest.

Validation of microbiome biomarkers. Microbiome biomarkers of a given condition are often strongly study dependent. It is thus crucial to validate biomarkers across technologies and cohorts to enhance reproducibility and minimize batch effects.

Data sharing and analysis reproducibility. Data and metadata sharing is strongly encouraged; raw data deposition is usually requested before publication and open-source software is desirable. However, metagenomics has yet to reach the level of standardization characteristic of other, more established high-throughput techniques.

trees, functional profiles and comparisons of MAG abundance across samples (**Fig. 2**, **Supplementary Code** and https://github.com/chris-quince/metag-rev-sup).

Assembly-free metagenomic profiling

Taxonomic profiling of metagenomes identifies which microbial species are present in a metagenome and estimates their abundance. This can be carried out without assembly, through external sequence data resources, such as publicly available reference genomes. This approach can mitigate assembly problems, speed up computation and enable profiling of low-abundance organisms that cannot be assembled *de novo* (**Supplementary Box 1**). Its main limitation is that previously uncharacterized microbes are difficult to profile (**Supplementary Box 1**). However, the number of reference genomes available is increasing rapidly, with thousands of genomes produced each year, including some derived from difficult-to-grow species targeted by new cultivation methods[70], single-cell sequencing approaches[71] or metagenomic assembly. The diversity of reference genomes available for some sample types, such as the human gut[72], is now extensive enough to make assembly-free taxonomic profiling efficient and successful, including for comparatively low-abundance microbes that lack sufficient sequence coverage and depth to enable assembly of their genome. Analysis of more diverse environments, including soil and oceans, is hampered by a lack of representative reference genomes. As a result, it is generally advisable to use assembly when analyzing metagenomes from these environments.

Assembly-free taxonomic profilers with species-level resolution utilize information available in reference genomes (N.S.)[73] and environment-specific assemblies[74] and have been used in the largest human-associated metagenomics investigations performed so far[2,5,74–79]. The simple brute-force mapping of reads to genomes can result in profiles with many false positives, but this approach has nonetheless been proven effective when the output is post-processed on the basis of lowest common ancestor (LCA) strategies[80] or coupled with compositional interpolated Markov models[81]. However, the run times of these approaches do not improve on those of assembly-based methods. Kraken[82] also exploits LCA but speeds up the computation by substituting sequence mapping with *k*-mer matching.

Taxonomic profiling by selecting representative or discriminative genes (markers) from available reference sequences is another fast and accurate assembly-free approach that has been implemented with several variations. By looking at co-abundant markers from preassembled environment-specific gene catalogs[83–85], for example, the MetaHIT consortium was able to characterize known and novel organisms in the human gut[5,74]. Similarly, mOTU[85] focuses on universally conserved but phylogenetically informative markers (e.g., genes encoding ribosomal proteins), whereas MetaPhlAn (N.S.)[86,87] (**Fig. 2**) adopts several thousand clade-specific markers with high discriminatory power and was effective to quantitatively profile the microbiome from multiple body areas for the Human Microbiome Project (HMP)[2] with a very low false positive discovery rate. These methods are scalable and can be used for large metagenomics meta-analyses (N.S.)[88]. Marker-based approaches can also be used for strain-level comparative microbial genomics using thousands of metagenomes (N.S.)[87–90]. Importantly, the accuracy of these methods will improve as more reference genomes and high-quality metagenomic assemblies become available. For large data sets with hundreds of samples on which performing or interpreting metagenomics assembly is impractical, marker-based approaches are currently the method of choice, especially for environments with a substantial fraction of microbial diversity covered by well-characterized sequenced species.

Genes and metabolic pathways from metagenomes

With a fragmented but high-quality metagenome assembly, the genetic repertoire of a microbial community can be identified using adapted single-genome characterization tools. These include a gene identification step, usually with a metagenomic-specific parameter setting[91], followed by homology-based annotation pipelines commonly used for characterizing pure isolate genome assemblies (**Fig. 2**). Indeed, some of the largest shotgun sequencing efforts so far[5] have used metagenomic assemblies to compile the microbial gene catalog of the human[92] and mouse[83] gut metagenomes, although this approach is often limited by the large fraction of uncharacterized genes in the reference database catalogs.

Other large metagenomic data sets[2] were interpreted by translated sequence searches against functionally characterized protein families (N.S.)[93]. Databases that include combinations of manually annotated and computationally predicted proteins families, such as KEGG[94] or UniProt[95], can be used for this task and enable characterization of the functional potential of the microbiome (**Fig. 2**). Single protein families are aggregated into higher-level metabolic pathways and functional modules, providing either graphical reports[80] or comprehensive metabolic presence, absence and abundance tables, as in the HUMAnN pipeline[93]. Regardless of whether an assembly-free or assembly-based approach is adopted, the main limiting factor in profiling the metabolic potential of a community is the lack of annotations for accessory genes in most microbial species (with the exception of selected model organisms; **Box 1**). This means that highly conserved pathways and housekeeping functions are more consistently detected and quantified in metagenomes, which might explain why functional traits are often surprisingly consistent across different samples and environments, even when taxonomic composition is highly variable[2]. Experimental characterization of microbial proteins, coding genes and other genomic features (tRNAs, noncoding RNAs and CRISPRS) to more thoroughly assess functions of individual loci is a bottleneck that currently has a crucial impact on the ability to profile the functions of metagenomes[84].

A complementary approach to metabolic function profiling of metagenomes is an in-depth characterization of specific functions of interest. For example, identification of genes involved in antibiotic resistance (the 'resistome') in a microbial community can inform on the spread of antibiotic resistance[96]. *Ad hoc* methods (N.S.)[97] and manually curated databases of antibiotic-resistance genes have been crucial to this approach; ARDB[98] was the first widely adopted resistance database and is now complemented by additional resources, such as Resfams[99]. Comparably large efforts are also devoted to reporting the virulence repertoire of a metagenome; targeted analyses of metagenomes for specific gene families of interest can also be used to validate findings from single, cultivation-based isolate experiments.

Post-processing analysis

Regardless of the methods used for primary metagenomic sequence analyses, the outputs will comprise data matrices of samples versus microbial features (i.e., species, taxa, genes and pathways). Post-processing analysis uses statistical tools to interpret these matrices and decipher how the findings correlate with the sample metadata. Many of these statistical approaches are not specific for metagenomics. Specific challenges of metagenome-derived quantitative values include the proportional nature of the taxonomic and functional profiles and the log-normal long-tailed distribution of abundances. These issues are also problematic in high-throughput 16S rRNA gene amplicon-sequencing data sets, and several popular R packages, such as DESeq2 (ref. 100), vegan[101] and metagenomeSeq[102], which were originally developed for amplicon sequencing, can be used for metagenomics.

Table 4 Strengths and weaknesses of assembly-based and read-based analyses for primary analysis of metagenomics data

	Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference databases.
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly.	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives.	Cannot resolve organisms for which genomes of close relatives are unknown.
Computational burden	Requires computationally costly assembly, mapping and binning.	Can be performed efficiently, enabling large meta-analyses.
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity.	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes.
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection.	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision.
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates.	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates.

Post-processing tools include traditional multivariate statistics and machine learning. Unsupervised methods include simple clustering and correlation of samples, and visualization techniques such as heat maps, ordination (e.g., principal component analysis and principal coordinates analysis) or networks, which allow patterns in the data to be revealed graphically. Some unsupervised statistical tools aim to specifically address the problems introduced by the proportional nature of metagenome profiles (compositionality issue)[103] (**Box 1**) and infer ecological relationships within the community (N.S.)[104]. Supervised methods include statistical methods, such as multivariate analysis of variance (ANOVA) for direct hypothesis testing of differences between groups, or machine learning classifiers that train models to label groups of samples, such as random forests or support vector machines (N.S.)[105]. A classic machine-learning example would be to diagnose disease (e.g., type 2 diabetes)[75] on the basis of community dysbiosis, although developing cross-study predictive signatures is challenging[105].

Unsupervised and supervised methods consider the community as a whole. A complementary strategy is to ask which specific taxa or functional genes are statistically different between sample types or patient groups. Given the complexity of metagenomics data sets and the huge numbers of comparisons that can typically be made, correction for multiple comparisons[106] or effect size estimation (N.S.)[107] are vital for this task.

Robust statistical testing is key to determining the validity of results, but compact graphical representations can intuitively reveal patterns. In many cases visualization of post-processing results requires *ad hoc* graphical tools[108,109] and carefully adopted general visualization approaches.

Outlook

Metagenomics still faces roadblocks to applicability, usefulness and standardization (**Box 1**). The lack of reference genome sequence data for large portions of the microbial tree of life and functional annotation for many microbial genes substantially reduces the potential for success of the computational approaches that are used to analyze the vast amounts of sequences produced. Metagenomes from environments such as soil or water are particularly affected by this problem owing to their high microbial diversity and the proportion of uncharacterized taxa in these communities. Shotgun sequencing also fails to discriminate between live and dead organisms. However, the outlook is bright, because a large community of wet-lab and computational researchers are gradually finding solutions to these problems.

Metagenome bioinformatics tools, especially for translating raw reads into meaningful microbial features (genomes, species abundances and functional potential profiles) (**Fig. 1**), are continually improving. For example, strain-level analyses are now possible[110–112]. There remains an active debate about which sequence analysis approach is best (**Table 4**). Metagenomic assembly is the preferred theoretical solution if there is sufficient genome coverage (i.e., more than 20×), but this level of coverage is difficult to obtain for most of the members of the microbiome (**Table 4**), and assembly-free methods have other advantages, including the potential to perform large-scale strain-level analyses. The success of either approach depends on the microbial community composition and complexity, sequencing depth, size of the data set and available computational resources (**Table 4**). We recommend that researchers use both approaches for sequence analysis whenever possible, as they complement and validate each other.

As for technological improvements in the sequencing of community DNA, long-read sequencing platforms have matured and are likely to become useful for metagenomics assembly strategies, although publications are few at present. Pacific Biosciences instruments can deliver complete or nearly complete isolated microbial genomes with low base-error rates if sufficient coverage is achieved (typically 30–100×). The Oxford Nanopore MinION, which is a single-molecule, long-read device, holds appeal because of its size and portability (comparable to a smartphone), and early analysis of reads from this platform indicates it has an error rate close to that of Pacific Biosciences reads (N.J.L.)[113]. Assembly of isolate genomes into single contigs is possible (J.S. and N.J.L.)[114], so the portability of the MinION raises the tantalizing possibility of metagenomic sequencing in the field.

An alternative experimental approach to improving genome reconstruction from metagenomes couples Illumina sequencing chemistry with a multiplexed pooling library preparation protocol. This 'synthetic long read' technology relies on the dilution of genomic DNA into fragmented and barcoded pools consisting of hundreds to thousands of individual molecules. These pools are sequenced and assembled *de novo* to produce synthetic long reads. One benefit of synthetic long reads is that because they are built from a consensus of Illumina sequences, the base error rate is extremely low. However, the protocol is rather laborious and requires high DNA input (between 1 and 10 μg), and problems persist with local repetitive sequences. Reports suggest that this approach is useful for metagenomics, especially when coupled with standard shotgun sequencing, as it can reconstruct genomes from closely related strains, as well as those from rare microorganisms[115,116].

Another outstanding problem in shotgun metagenomics is the accurate reconstruction of strain-level variation from mixtures of genetically related organisms[117], with several proposed solutions (C.Q. and N.S.)[14,89,110–113,118,119] based on assembly, mapping or both. Mapping to genes that are unique to a species (N.S.)[87] can resolve the dominant haplotype in a sample, and this method has been applied to thousands of unrelated metagenomes, providing strain-level phylogenies that enable microbial population genomics for hundreds of largely uncharacterized species (N.S.)[110]. Mixtures of strains from the same species in a single sample cannot be resolved by consensus approaches, but if the same strains are present in multiple samples, there will be characteristic signatures in single-nucleotide variations. These nucleotide variations can be linked to deduce haplotypes and their frequencies (C.Q.)[89,112,113,118]. This methodology was initially applied only after mapping to reference genes[89] and, optionally, with simultaneous strain phylogeny reconstruction[118], but it has now been applied directly to contig bins with inference of strain gene complement in an entirely reference-free method (C.Q.)[112]. One limitation of this approach is that in some environments, including the human gut, one strain usually dominates over other strains from the same species (N.S.)[110]. It is therefore challenging to detect nondominant strains of low-abundance species, and the user has to weigh the increased robustness of profiling only the dominant strains (N.S.)[110] with the potential additional information that can be garnered from characterizing mixtures of strains (C.Q.)[112]. Strain-level metagenomics is a very active area of research[117] and has the potential to empower metagenomics with a resolution similar to that derived from sequencing cultured single isolates. Although long-read technologies could aid these efforts in the future, until then, solving the computational challenges of strain-level profiling from metagenomics is arguably the biggest challenge the field faces.

Conclusions

Since the pioneering application of whole DNA sequencing to environmental samples by teams led by Jillian Banfield[120] and J. Craig Venter[7] in 2004, shotgun metagenomics has become an important tool for the study of microbial communities. Widespread adoption of metagenomics has been enabled by the falling cost of sequencing and the development of improved computational methods. The main limitations now facing researchers are the costs of training computational scientists to analyze the complex metagenomic data sets and of sequencing enough samples for properly powered study designs. Initiatives such as the Critical Assessment of Metagenomic Interpretation (CAMI) (C.Q.)[55] are vital for an unbiased assessment of computational tools to improve reproducibility and standardization.

Shotgun metagenomics will have an increasingly important part in diverse biomedical and environmental applications. We hope this review will provide an understanding of the basic concepts of shotgun metagenomics, including its limitations and its immense potential.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

A.W.W. and the Rowett Institute receive core funding support from the Scottish Government's Rural and Environmental Science and Analysis Service (RESAS). N.S. is supported by the European Research Council (ERC-STG project MetaPG), a European Union Framework Program 7 Marie-Curie grant (PCIG13-618833), a MIUR grant (FIR RBFR13EWWI), a Fondazione Caritro grant (Rif.Int.2013.0239) and a Terme di Comano grant. C.Q. and N.J.L. are funded through a MRC bioinformatics fellowship (MR/M50161X/1) as part of the MRC Cloud Infrastructure for Microbial Bioinformatics (CLIMB) consortium (MR/L015080/1). J.T.S. is supported by the Ontario Institute for Cancer Research through funding provided by the Government of Ontario.

AUTHOR CONTRIBUTIONS

C.Q., A.W.W., J.T.S., N.J.L. and N.S. drafted the paper, revised the text and designed figures, tables and boxes. C.Q. and N.S. performed the metagenomic analyses described in the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* **19**, 1141–1152 (2009).
2. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
3. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).
4. Loman, N.J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *J. Am. Med. Assoc.* **309**, 1502–1510 (2013).
5. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
6. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
7. Venter, J.C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
8. Brown, C.T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
9. van Kessel, M.A. *et al.* Complete nitrification by a single microorganism. *Nature* **528**, 555–559 (2015).
10. Daims, H. *et al.* Complete nitrification by *Nitrospira* bacteria. *Nature* **528**, 504–509 (2015).
11. Donia, M.S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).
12. Norman, J.M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).
13. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
14. Li, S.S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586–589 (2016).
15. Kuczynski, J. *et al.* Direct sequencing of the human microbiome readily reveals community differences. *Genome Biol.* **11**, 210 (2010).
16. Goodrich, J.K. *et al.* Conducting a microbiome study. *Cell* **158**, 250–262 (2014).
17. La Rosa, P.S. *et al.* Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One* **7**, e52078 (2012).
18. Tickle, T.L., Segata, N., Waldron, L., Weingart, U. & Huttenhower, C. Two-stage microbial community experimental design. *ISME J.* **7**, 2330–2339 (2013).
19. Bonder, M.J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
20. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
21. Knight, R. *et al.* Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* **30**, 513–520 (2012).
22. McCafferty, J. *et al.* Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. *ISME J.* **7**, 2116–2125 (2013).
23. Lees, H. *et al.* Age and microenvironment outweigh genetic influence on the Zucker rat microbiome. *PLoS One* **9**, e100916 (2014).
24. Sorge, R.E. *et al.* Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* **11**, 629–632 (2014).
25. Laukens, D., Brinkman, B.M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS Microbiol. Rev.* **40**, 117–132 (2016).
26. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
27. Lozupone, C.A. *et al.* Meta-analyses of studies of the human microbiota. *Genome Res.* **23**, 1704–1714 (2013).
28. Probst, A.J., Weinmaier, T., DeSantis, T.Z., Santo Domingo, J.W. & Ashbolt, N. New perspectives on microbial community distortion after whole-genome amplification. *PLoS One* **10**, e0124158 (2015).

29. Cuthbertson, L. *et al.* Time between collection and storage significantly influences bacterial sequence composition in sputum samples from cystic fibrosis respiratory infections. *J. Clin. Microbiol.* **52**, 3011–3016 (2014).
30. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19 (2014).
31. Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z. & Forney, L.J. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* **7**, e33865 (2012).
32. Kennedy, N.A. *et al.* The impact of different DNA extraction kits and laboratories upon the assessment of human gut microbiota composition by 16S rRNA gene sequencing. *PLoS One* **9**, e88982 (2014).
33. Tanner, M.A., Goebel, B.M., Dojka, M.A. & Pace, N.R. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl. Environ. Microbiol.* **64**, 3110–3113 (1998).
34. Salter, S.J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
35. Motley, S.T. *et al.* Improved multiple displacement amplification (iMDA) and ultraclean reagents. *BMC Genomics* **15**, 443 (2014).
36. Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L. & Graf, J. Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One* **9**, e94249 (2014).
37. Sinha, R. *et al.* Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. Preprint at http://www.biorxiv.org/content/early/2017/04/09/125724 (2017).
38. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036 (2015).
39. Jones, M.B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc. Natl. Acad. Sci. USA* **112**, 14024–14029 (2015).
40. Simpson, J.T. & Pop, M. The theory and practice of genome sequence assembly. *Annu. Rev. Genomics Hum. Genet.* **16**, 153–172 (2015).
41. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
42. Simpson, J.T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**, 1228–1235 (2014).
43. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. *De novo* assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).
44. Peng, Y., Leung, H.C., Yiu, S.M. & Chin, F.Y. Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* **27**, i94–i101 (2011).
45. Namiki, T., Hachiya, T., Tanaka, H. & Sakakibara, Y. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* **40**, e155 (2012).
46. Peng, Y., Leung, H.C., Yiu, S.M. & Chin, F.Y. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
47. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
48. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
49. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biol.* **13**, R122 (2012).
50. Pell, J. *et al.* Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc. Natl. Acad. Sci. USA* **109**, 13272–13277 (2012).
51. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
52. Li, D., Liu, C.M., Luo, R., Sadakane, K. & Lam, T.W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
53. Bradnam, K.R. *et al.* Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
54. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55 (2016).
55. Sczyrba, A. *et al.* Critical assessment of metagenome interpretation—a benchmark of computational metagenomics software. Preprint at http://www.biorxiv.org/content/early/2017/06/12/099127 (2017).
56. Karlin, S., Mrázek, J. & Campbell, A.M. Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.* **179**, 3899–3913 (1997).
57. Dick, G.J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10**, R85 (2009).
58. Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. & Sokhansanj, B. Metagenome fragment classification using *N*-mer frequency profiles. *Adv. Bioinformatics* **2008**, 205969 (2008).
59. McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* **4**, 63–72 (2007).
60. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
61. Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H.E. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* **3**, 410 (2012).
62. Kelley, D.R. & Salzberg, S.L. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics* **11**, 544 (2010).
63. 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).
64. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
65. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–1106 (2015).
66. Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).
67. Kang, D.D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
68. Eren, A.M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
69. Hug, L.A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
70. Stewart, E.J. Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151–4160 (2012).
71. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
72. Nelson, K.E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
73. Segata, N. *et al.* Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666 (2013).
74. Nielsen, H.B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
75. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
76. Karlsson, F.H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
77. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
78. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
79. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
80. Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S.C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560 (2011).
81. Brady, A. & Salzberg, S.L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**, 673–676 (2009).
82. Wood, D.E. & Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
83. Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat. Biotechnol.* **33**, 1103–1108 (2015).
84. Walker, A.W., Duncan, S.H., Louis, P. & Flint, H.J. Phylogeny, culturing, and metagenomics of the human gut microbiota. *Trends Microbiol.* **22**, 267–274 (2014).
85. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
86. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
87. Truong, D.T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
88. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* (in press).
89. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* **33**, 1045–1052 (2015).
90. Donati, C. *et al.* Uncovering oral *Neisseria* tropism and persistence using metagenomic sequencing. *Nat. Microbiol.* **1**, 16070 (2016).
91. Zhu, W., Lomsadze, A. & Borodovsky, M. *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).
92. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
93. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
94. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
95. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42**, D191–D198 (2014).
96. Pehrsson, E.C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212–216 (2016).
97. Kaminski, J. *et al.* High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.* **11**, e1004557 (2015).

98. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37**, D443–D447 (2009).
99. Gibson, M.K., Forsberg, K.J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
100. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 0550–8 (2014).
101. Oksanen, J. *et al.* Vegan: the community ecology package. *The Comprehensive R Archive Network* https://cran.r-project.org/web/packages/vegan/vegan.pdf (2007).
102. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10**, 1200–1202 (2013).
103. Friedman, J. & Alm, E.J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).
104. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8**, e1002606 (2012).
105. Pasolli, E., Truong, D.T., Malik, F., Waldron, L. & Segata, N. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* **12**, e1004977 (2016).
106. White, J.R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* **5**, e1000352 (2009).
107. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).
108. Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C. & Segata, N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* **3**, e1029 (2015).
109. Ondov, B.D., Bergman, N.H. & Phillippy, A.M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).
110. Duy Truong, T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* **27**, 626–638 (2017).
111. Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* **13**, 435–438 (2016).
112. Quince, C. *et al. De novo* extraction of microbial strains from metagenomes reveals intra-species niche partitioning. Preprint at http://www.biorxiv.org/content/early/2016/09/06/073825 (2016).
113. Quick, J., Quinlan, A.R. & Loman, N.J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* **3**, 22 (2014).
114. Loman, N.J., Quick, J. & Simpson, J.T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
115. Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat. Biotechnol.* **34**, 64–69 (2016).
116. Sharon, I. *et al.* Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Res.* **25**, 534–543 (2015).
117. Marx, V. Microbiology: the road to strain-level identification. *Nat. Methods* **13**, 401–404 (2016).
118. O'Brien, J.D. *et al.* A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics* **197**, 925–937 (2014).
119. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K.S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).
120. Tyson, G.W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
121. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
122. de Bourcy, C.F. *et al.* A quantitative comparison of single-cell whole-genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
123. Yilmaz, S., Haroon, M.F., Rabkin, B.A., Tyson, G.W. & Hugenholtz, P. Fixation-free fluorescence *in situ* hybridization for targeted enrichment of microbial populations. *ISME J.* **4**, 1352–1356 (2010).
124. Delmont, T.O. *et al.* Reconstructing rare soil microbial genomes using *in situ* enrichments and metagenomics. *Front. Microbiol.* **6**, 358 (2015).
125. Kent, B.N. *et al.* Complete bacteriophage transfer in a bacterial endosymbiont (*Wolbachia*) determined by targeted genome capture. *Genome Biol. Evol.* **3**, 209–218 (2011).
126. Seth-Smith, H.M. *et al.* Generating whole bacterial genome sequences of low-abundance species from complex samples with IMS-MDA. *Nat. Protoc.* **8**, 2404–2412 (2013).
127. Lim, Y.W. *et al.* Purifying the impure: sequencing metagenomes and metatranscriptomes from complex animal-associated samples. *J. Vis. Exp.* **94**, e52117 (2014).
128. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
129. Ofek-Lalzar, M. *et al.* Niche and host-associated functional signatures of the root surface microbiome. *Nat. Commun.* **5**, 4950 (2014).
130. Hyatt, D., LoCascio, P.F., Hauser, L.J. & Uberbacher, E.C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
131. Huson, D.H. *et al.* Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. *Microbiome* **5**, 11 (2017).
132. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
133. Marchler-Bauer, A. *et al.* CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30**, 281–283 (2002).
134. Beszteri, B., Temperton, B., Frickenhaus, S. & Giovannoni, S.J. Average genome size: a potential source of bias in comparative metagenomics. *ISME J.* **4**, 1075–1077 (2010).
135. Knight, R. *et al.* Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* **30**, 513–520 (2012).

Corrigendum: Shotgun metagenomics, from sampling to analysis

Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman & Nicola Segata
***Nat. Biotechnol.* 35, 833–844 (2017); published online 12 September 2017; corrected after print 12 September 2017**

In the version of this article initially published, the Competing Financial Interests should have indicated the authors had competing interests, but instead indicated there were none. The detailed statement was missing from the HTML: J.T.S. receives research funding from Oxford Nanopore Technologies and has received travel and accommodations to speak at meetings hosted by Oxford Nanopore Technologies. N.J.L. has received honoraria to speak at Oxford Nanopore and Illumina meetings, and travel and accommodation to attend company-sponsored meetings. N.J.L. has ongoing research collaborations with Oxford Nanopore who have provided free-of-charge sequencing reagents as part of the MinION Access Programme and directly in support of research projects. In addition, the publication date was given as 11 September, rather than 12 September 2017. The errors have been corrected for the PDF and HTML versions of this article.