Shotgun metagenomics, from sampling to analysis.

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

[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, Canada
[4] Department of Computer Science, University of Toronto, Toronto, Canada.
[5] Institute for Microbiology and Infection, University of Birmingham, Birmingham, UK.
[6] Centre for Integrative Biology, University of Trento, Trento, Italy.

^ These authors contributed equally
* Corresponding author: Nicola Segata (nicola.segata@unitn.it)

Editors summary

The promises and potential pitfalls of shotgun metagenomics, from experimental design to computational analyses, are reviewed.

General:

Please ensure any self-cites in the text are identified. This should be done for any references that any of the 5 authors are involved in, with this format: text text text [21] ((C.Q., N.J.L)

Please ensure any bioRXiv references are updated should the papers now be published

Display items

There are four figures, three tables, and three 'BOXES' (at least two of which look like tables). This is too many elements for one article and we need to prune this to 7 display items in total (you have 10). I suggest the following – remove figure 1, figure 2. Retain figure 3 but simplify it (see annotations), retain figure 4 but simplify (see annotations). BOX 1 repeats the text, delete it. Box 2 is a table, change it into a table. Box 3 needs to be made into prose, its formatted as a table.

OUR SUGGESTED UPDATED PLAN

Original Fig 1 → revised **Fig 1** (display item 1)

Original Fig 2 → Supplementary Fig 1

Original Fig 3 → revised **Fig 2** (display item 2)

Original Fig 4 → Deleted

Original Table 1 → Revised **Table 1** (display item 3)

Original Table 2 → Revised **Table 2** (display item 4)

Original Table 3 → Revised **Table 3** (display item 5)

Original Box 1 → Supplementary Box 1

Original Box 2 → Revised **Table 4** (display item 6)

Original Box 3 → **Box 1** (display item 7)

Abstract

Diverse microbial communities of bacteria, archaea, viruses and single-celled eukaryotes have crucial roles in the environment and human health. However, microbes are frequently difficult to culture in the laboratory, which can confound cataloging members and understanding how communities function. Cheap, 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 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 offer biotechnological promise in therapeutic discovery, or innovative ways to synthesize products using microbial factories, but can also pinpoint the contributions of microorganisms to planetary, animal and human health.

Introduction

High throughput sequencing approaches enable genomic analyses of ideally all microbes in a sample, not just those that are more amenable to cultivation. One such method, shotgun metagenomics, is the untargeted ("shotgun") sequencing of all ("meta") of the 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 mistakenly referred to as metagenomics but are not metagenomic methods, because they do not target the entire genomic content of a sample.

In the past 15 years since it was first used, metagenomics has enabled large-scale investigations of complex microbiomes[2-7](ref#4 C.Q.,N.J.L.). Discoveries enabled by this technology include the identification of previously unknown environmental bacterial phyla with endosymbiotic behavior [8], and species that can carry out complete nitrification of ammonia [9,10]. Other striking findings include the widespread presence of antibiotic genes in commensal gut bacteria [11], tracking of human outbreak pathogens [4](C.Q.,N.J.L.), the strong association of both the viral [12] and bacterial [13] fraction 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 faecal microbiome transplantation [14].

In this Review we discuss best-practice for shotgun metagenomics studies, including identifying and tackling limitations, and provide an outlook for metagenomics in the future.

Shotgun metagenomics study design

A typical shotgun metagenomics study comprises five steps following the initial study design; (i) the collection, processing, and sequencing of the samples, (ii) the preprocessing of the sequencing reads, (iii) the sequence analysis to profile taxonomic, functional, and genomic features of the microbiome, (iv) the postprocessing statistical and biological analysis, and (v) the validation (**Figure 1**). Numerous experimental and computational approaches are available to carry out each step, which means that researchers are faced with a daunting choice. And, despite its apparent simplicity, shotgun metagenomics has limitations, owing to potential experimental biases and the complexity of

computational analysis and their interpretation. We assess the choices that need to be made at each step and how to overcome common problems.

The steps involved in the design of hypothesis-based studies are outlined in **Supplementary Figure 1** with specific recommendations summarized in **Supplementary Box 1**. Individual samples from the same environment can be variable in microbial content, which makes it challenging to detect 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-tier 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 [18](N.S).

Controls can be difficult to obtain, particularly for samples from complex environments. This is particularly important for those studying 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 not rely on results from a single sample that might be a non-representative 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 impact the microbiome, e.g. antibiotics. If this is not feasible, then potential confounders should be factored into comparative analyses (see **Supplementary Box 1**).

If samples originate in animal models, particularly those involving co-housed rodents, the roles of animal age and housing environment [22,23], and the sex of the person handling the animals [24], may have on microbial community profiles should be taken into account. It is usually possible to mitigate against potential confounders in the study design by housing animals individually to prevent the spread of microbes between cage mates (although this may introduce behavioural changes, potentially resulting in different biases), mixing animals derived from different experimental cohorts together within the same cage, or repeating experiments with mouse lines obtained 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. MiMARKS and MIxS standards were set out to provide guidance for required metadata [26], but metagenomics is now applied on such disparate kinds 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, in order 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 meta-analyses of datasets from different studies (**Supplementary Box 1**). Collection and storage methods that have been validated for one type of sample type cannot be assumed to be optimal for different sample types. As such, careful preliminary work to optimize processing conditions for sample types is often necessary (**Supplementary Figure 1**).

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

The choice of DNA extraction method can affect the composition of downstream sequence data [30]. The extraction method must be able to lyse diverse microbial taxa, otherwise sequencing results may be dominated by DNA derived 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, bead-beating based approaches do vary in their efficiency [32] (A.W.W). 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 be during sample processing stages. Kit/laboratory reagents may contain variable amounts of microbial contaminants [33]. Metagenomics datasets from low biomass samples (e.g. skin swabs) are particularly vulnerable to this problem, because there is less “real” signal to compete with low-levels of contamination [34] (A.W.W.,N.J.L). We advise those 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 [34] (A.W.W.,N.J.L). Other types of contamination are cross-over from previous sequencing runs, presence of PhiX control DNA that is typically used as part of Illumina-based sequencing protocols, and human or host DNA.

Library preparation and sequencing

Choosing a library preparation and sequencing method hinges on availability of materials and services, cost, ease of automation, and DNA sample quantification. The Illumina platform has become dominant as a choice for shotgun metagenomics due to its wide availability, very high outputs (up to 1.5 Tb per run) and high accuracy (with a typical error rate of between 0.1-1%), although the competing Ion Torrent S5/S5 XL instrument is an alternative choice. Recently, long read sequencing technologies such as the Oxford Nanopore MinION and Pacific Biosciences Sequel have scaled up output and can reliably generate up to 10 gigabases per run and may therefore soon start to see adoption for metagenomics studies.

Given the very high outputs achievable on a single instrument run, multiple metagenomic samples are usually sequenced on the same sequencing run, by multiplexing up to 96 or 384 samples typically using dual indexing barcode sets available for all library preparation protocols. The Illumina platforms are known to suffer from issues of carry-over (between runs) and carry-between (within runs) [36]. Recently, concern has been raised that newer Illumina instruments using isothermal cluster generation (ExAmp) suffer from high rates of 'index hopping' where incorrect barcode identifiers are incorporated into growing clusters [37] although the extent of this problem on typical metagenomics projects has not been evaluated and approaches to mitigate it have been suggested. To help evaluate the extent of such issues, randomly chosen control wells containing known spiked-in organisms as positive controls, and template negative controlsshould be used to assess the impact of these issues. Such controls 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, performing technical replicates would be useful to assess variability, and even subjecting 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", for example in the Illumina Nextera and Nextera XT products, are popular owing to their low cost (list prices of $25-40 per sample, with dilution methods potentially able to reduce these costs even further [38]). Tagmentation approaches only require small DNA inputs (1 ng of DNA recommended, but lower amounts can be used). Such low inputs are achieved due 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 of reducing 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 therefore 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 2500/4000, NextSeq, and NovaSeq all produce high volumes of sequence data (between 120 gigabases and 1.5 terabases per run) and are well suited for metagenomics studies (with the caveat of index hopping). The throughput per run of these instruments is known and, by deciding the level of multiplexing, the investigator can set the desired per-sample sequencing depth. Typical experiments in 2017 aim to generate between 1 and 10 gigabases, but these depths may be either excessive or woefully little depending on the sensitivity required to detect rare members of a sample.

The Illumina platforms mainly differ by 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 2x250 nt in rapid run mode (generating up to 180 Gb per flowcell), or up to 1Tb in high output mode with 2x125 nt reads. The newer HiSeq 3000 and 4000 systems further increase the overall throughput of a run (up to 1.5 terabases for the 4000) but are limited to read lengths of 150 nt.

The NextSeq benchtop instrument has similar output to the HiSeq 2500's rapid run mode, but are limited to 150 nt reads. However the NextSeq is less than half the price of the HiSeq and so may be attractive to research groups wishing to operate their own instrument The recently released NovaSeq platform promises up to 3 terabases per run in the near future. The Illumina MiSeq is limited by output (up to 15Gb in 2x300 mode) but remains the *de facto* standard for single marker gene microbiome studies. The MiSeq (or MiniSeq) may still be useful for metagenomics 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 (and much more expensive) instruments where individual runs may cost >$10,000.

Metagenome assembly

Numerous approaches to computationally reconstruct the composition of the microbial community from the pool of sequence reads have been published. Choosing the "best" approach is a daunting task but largely depends on the aims of the study.

Metagenome *de novo* assembly, is conceptually similar to whole genome assembly [40](J.S.). The de Bruijn graph approach [41] is currently the most 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 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 non-genomic sequences that must be avoided, and repetitive sequence, which can cause misassemblies and fragmentation of the assembly.

Metagenome assembly presents challenges not faced in single genome assembly. 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 [42](J.S.) 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 increasing the 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 differ by a single nucleotide variant, or by the presence/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 difficult 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 [48](J.S.) such as Ray, which spread memory load over a cluster of computers, have been used to assemble metagenomes from human faecal samples [49]. To help assemble very complex samples Pell *et al*. developed a lightweight method to partition a metagenome assembly graph into connected components that can be assembled independently [50]. Another method, named 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 will depend on biological factors such as underlying microbial community structure, and technical factors, such as sequencing platform characteristics and coverage. This effect was observed at an Assemblathon [53], where no single assembler came out "best".

We analysed assembly results from mock synthetic and real communities (**Table 2** and **Table 3)**. We evaluated two assemblers, MegaHIT [52] and MetaSPAdes [54] for their ability to reconstruct known genomes from the mock communities, and capture taxonomic and gene diversity in the real datasets. They both successfully reconstructed more than 75% of the mock communities (one comprising 20 organisms [2], the other 49 bacterial and 10 archaeal species [55](C.Q.)). MetaSPAdes generated longer contigs, but these appeared to be less accurate. When restricted to contigs that exactly matched the references in the mock community then 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 datasets (**Table 3**), consistent patterns were hard to discern. However, examining median single-copy core gene number (which will estimate the number of genomes in the assembly) suggests that for the more complex soil and ocean communities, MegaHIT succeeded in assembling more genes that could then be functionally annotated. However, the key message here is that different state-of-the-art programs will be optimal on different datasets while requiring similar run times (about 48 hours using 16 threads on the largest sample) and main memory usage (not exceeding 125GB). It is prudent, therefore, to attempt more than one assembly approach. The CAMI challenge reported that MegaHIT was in the top three best metagenomics assemblers across their benchmark data sets [56](C.Q.) and together with MetaSPAdes (not evaluated in CAMI) these are probably the best current choices. Whatever assembler is used the result will not be genomes but rather potentially millions of contigs, and this motivates the need for binners that attempt to link those contigs back into the genomes they derived from.

Binning contigs

Metagenome assemblies are highly fragmented, comprising thousands of contigs (**Table 2**), and the challenge is that we do not know *a priori* which contig derives from which genome. We do not 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 methods, or clustering, 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 [57]. Metrics based on these k-mer frequencies can be used to bin contigs, with tetramers considered the most informative for binning of metagenomics data [58]. Many different software choices are available that are based on these frequencies such as Naïve Bayes classifiers [59] or support vector machines [60], but sequence composition often lacks the specificity necessary to resolve complex datasets to the species level in complex communities [58,61](ref#61 C.Q.,N.J.L.).

Clustering of contigs is appealing because it does not require reference genomes. Until recently, most contig clustering algorithms such as MetaWatt [62] and SCIMM [63] used various species composition metrics, sometimes coupled with total coverage. Recently, as multi-sample metagenome datasets have been produced it has been realized that contig coverage across multiple samples provides a much more powerful signal to group contigs together [64,65]. The 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 [66]. The first algorithms, e.g. extended self-organising maps [64], required human input to perform the clustering, which is based on coverage information and composition that could be visualized in 2D [65]. Completely automated approaches such as CONCOCT[61](C.Q.,N.J.L.), GroopM [67] and MetaBAT [68] are now available and they are convenient, particularly for large datasets, but better results may still be obtained when combined with human refinement, for instance using a visualization tool named Anvio [69](C.Q.).

Methods for reconstructing metagenomic assembled genomes (MAGs) are indispensable to uncover the hitherto inaccessible diversity of bacteria. The recovery of nearly a thousand 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 all small, with minimal metabolism, and formed a monophyletic clade, separate from the previously cultured diversity of bacteria. These have been proposed as a new bacterial sub-division, the candidate phyla radiation, revealed through metagenomics [70].

Completeness of MAGs is usually evaluated by examining single-copy core genes, which are found in most microbial genomes, for example tRNA synthetases or ribosomal proteins. A pure MAG will

have all these genes present in single copies. Once constructed, the MAGs provide a rich dataset for comparative genomics, including the construction of phylogenetic trees, functional profiles and comparisons of MAG abundance across samples (see left panel in **Figure 2** and the step-by-step tutorial we provide at https://github.com/chrisquince/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 using external sequence data resources, such as publicly available reference genomes. This approach can mitigate assembly problems, speed up computation, and make it possible to profile low-abundance organisms that cannot be assembled *de novo* (**Supplementary Box 1**). The main limitation is that previously uncharacterized microbes are very difficult to profile (**Supplementary Box 1**). However, the number of reference genomes available is increasing rapidly, with thousands of genomes being produced each year, including some derived from difficult-to-grow species targeted by new cultivation methods [71], single-cell sequencing approaches [72], or metagenomic assembly itself. The diversity of reference genomes available for some sample types, such as from the human gut [73], 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 the 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 inadvisable to avoid assembly when analyzing metagenomes from these environments.

Assembly-free taxonomic profilers with species-level resolution utilize information available in reference genomes [74](N.S) and in environment-specific assemblies [75], and have been used in the largest human-associated metagenomics investigations performed so far [2,5,75-80]. The simple brute force mapping of reads to genomes can result in profiles with many false positives but, nonetheless, this approach has been proven to be effective when the output is post-processed based on lowest common ancestor (LCA) strategies [81] or coupled with compositional interpolated Markov models [82]. However, the run times of these approaches do not improve on assembly-based methods. Kraken [83] also exploits LCA but dramatically 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 pre-assembled environment specific gene catalogs [84,85](ref#85 A.W.W.), for example, the MetaHIT consortium was able to characterize known and novel organisms in the human gut [5,75]. Similarly, mOTU [86] focuses on universally conserved but phylogenetically informative markers (e.g. genes coding for ribosomal proteins), whereas MetaPhlAn [87,88](N.S) (right panel of **Figure 2**) adopts several thousands of clade-specific markers with high discriminatory power, and proved effective to quantitatively profile the microbiome from multiple body areas for the Human Microbiome Project [2] with a very low false positive discovery rate. These methods are scalable and can be used for large metagenomics meta-analyses [89](N.S.). Marker-based approaches can also be used for strain-level comparative microbial genomics using thousands of

metagenomes [88,90,91](ref#88 N.S.). Importantly, the accuracy of these methods will improve as more reference genomes and high-quality metagenomic assemblies become available. For large datasets 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 gene repertoire of a microbial community can be identified using adaptations of single-genome characterization tools. These include a gene identification step, usually with a metagenomic-specific parameter setting [92], followed by homology-based annotation pipelines commonly used for characterizing pure isolate genome assemblies. Indeed, some of the largest shotgun sequencing efforts performed so far [5] used metagenomic assemblies to compile the microbial gene catalog of the human [93] and mouse [84] gut metagenomes, although this approach is often limited by the large fraction of uncharacterized genes in the reference database catalogs.

Other large metagenomic datasets [2] were interpreted by translated sequence searches against functionally characterized protein families [94](N.S.). Databases, that include combinations of manually annotated and computationally predicted proteins families such as KEGG [95] or UniProt [96], can be used for this task and enable characterization of the functional potential of the microbiome (**Figure 2, right-hand panel**). Single protein families are aggregated into higher-level metabolic pathways and functional modules providing either graphical reports [81] or comprehensive metabolic presence/absence and abundance tables, as in the HUMAnN pipeline [94](N.S.). 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 reported to be 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, non-coding RNAs, CRISPRS) to more thoroughly assess functions of individual loci is a bottleneck that currently has a crucial impact on our ability to profile the functions of metagenomes [85].

A complementary approach to metabolic function profiling of metagenomes is an in-depth characterization of specific functions of interest. For example, identifying genes involved in antibiotic resistance (the “resistome”) in a microbial community can inform on the spread of antibiotic resistance [97]. *Ad-hoc* methods [98](N.S.) and manually curated databases of antibiotic resistance genes have been crucial to this approach; ARDB [99] was the first widely adopted resistance database and is now complemented by additional resources such as Resfams [100]. 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 (species, taxa, genes, pathways). Post-processing analysis uses statistical tools to interpret these matrices, and decipher how the findings correlate with the sample meta-data. 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 datasets, and several popular R packages such as DESeq2 [101], vegan [102], and metagenomeSeq [103] that were originally developed for amplicon sequencing can be used for metagenomics.

Post-processing tools include traditional multivariate statistics and machine learning. Unsupervised methods include simple clustering and correlation of samples, andvisualization techniques such as heatmaps, ordination (e.g. PCA and PCoA), or networks, which allow the 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 [104], **Box 1**) and try to infer ecological relationships within the community [105](N.S.). Supervised methods include both statistical methods such as multivariate analysis of variance ANOVAs 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 [106](N.S.). A classic machine learning example would be to diagnose disease (e.g. for type 2 diabetes [76]) on the basis of community dysbiosis, although developing cross-study predictive signatures is challenging [106](N.S.).

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 datasets, and the huge numbers of comparisons that can typically be made, correction for multiple comparisons [107] or effect size estimation [108](N.S.) 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 [109,110](ref#109 N.S.), 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, or functional annotation for many microbial genes, substantially reduce the potential for success of the computational approaches used to analyse the vast amounts of sequences produced. Metagenomes from environments such as soil or water are particularly affected by this problem owing to both 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 year on year a large community of wet-lab and computational researchers are finding solutions to these problems.

Metagenome bioinformatics tools, especially for translating raw reads into meaningful microbial features (genomes, species abundances, functional potential profiles) (**Figure 1**), are continually improving. For example, strain-level analyses are now possible [111-113](ref#113 C.Q, ref#111 N.S., ref#112 N.S.). There remains an active debate about which sequence analysis approach is best (see **Table 4**). Metagenomic assembly is the preferred theoretical solution if there is sufficient genome coverage (i.e. >20x), 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 possibility 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 dataset, 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 the 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. The Pacific Biosciences instruments can deliver complete or nearly complete isolated microbial genomes with low base error rates if sufficient coverage is achieved (typically 30-100X). The Oxford Nanopore MinION single molecule, long read instrument holds appeal because of its size and portability (smartphone size) and early analysis of reads from this platform indicates it has an error rate akin to Pacific Biosciences reads [114](N.J.L.). Assembly of isolate genomes is possible into single contigs [115](J.S.,N.J.L.) so the portability of the MinION raises the tantalizing possibility of performing metagenomic sequencing in the field.

An alternative experimental approach to improve genome reconstruction from metagenomes couples Illumina sequencing chemistry with a multiplexed pooling library preparation protocol. This so-called Synthetic Long Reads 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 of DNA), plus, 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 [116,117].

Another outstanding problem in shotgun metagenomics is the accurate reconstruction of strain-level variation from mixtures of genetically related organisms [118], with several solutions proposed [14,90,111-113,119,120](ref#113 C.Q., ref#111 N.S., ref#112 N.S.) that are based on assembly, mapping, or a combination of the two. Mapping to genes that are unique to a species [88](N.S) 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 [111](N.S). 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

485 characteristic signatures in single nucleotide variations. These nucleotide variations can be linked
486 together to deduce haplotypes and their frequencies [90,113,119](ref#113 C.Q.). This methodology was
487 initially only applied after mapping to reference genes [90], and optionally with simultaneous strain
488 phylogeny reconstruction [119], but it has now been applied directly to contig bins with inference of strain
489 gene complement in an entirely reference free method [113](C.Q.). One limitation of this approach is that
490 in some environments, including the human gut, it has been shown that one strain usually dominates
491 over other strains from the same species [111](N.S). It is therefore challenging to detect non dominant
492 strains of low-abundance species, and the user has to weight the increased robustness of profiling only
493 the dominant strains [111](N.S) with the potential additional information that can be garnered from
494 characterizing mixtures of strains [113](C.Q.). Strain-level metagenomics is an active area of research [118]
495 and has the potential to empower metagenomics with similar resolution to that which can be derived
496 from sequencing of pure culture single isolates. Although long read technologies can aid these efforts in
497 the future, solving the computational challenges of strain-level profiling from metagenomics is arguably
498 the biggest challenge in the field at the moment.

499 ## Conclusions

500 Since the pioneering application of whole DNA sequencing to environmental samples by teams
501 led by Jillian Banfield [121] and Craig Venter [7] in 2004, shotgun metagenomics has become an important
502 tool for the study of microbial communities. Widespread adoption of metagenomics has been enabled
503 by the falling cost of sequencing and the development of tractable computational methods. The main
504 limitations facing researchers now are the costs of training computational scientists for analyzing the
505 complex metagenomic datasets, and of sequencing enough samples for properly powered study
506 designs. Initiatives such as the Critical Assessment of Metagenomic Interpretation [56](C.Q.) are vital for an
507 unbiased assessment of computational tools to improve reproducibility and standardization.

508 Shotgun metagenomics will play an increasingly important part in diverse biomedical and
509 environmental investigations and applications. We hope that this Review will provide an understanding
510 of the basic concepts of shotgun metagenomics including both its limitations and its immense potential.

511

512 ## Acknowledgments

513 AWW and The Rowett Institute, University of Aberdeen, receive core funding support from the Scottish
514 Government's Rural and Environmental Science and Analysis Service (RESAS). NS is supported by the
515 European Research Council (ERC-STG project MetaPG), European Union FP7 Marie-Curie grant (PCIG13-
516 618833), MIUR grant FIR RBFR13EWWI, Fondazione Caritro grant Rif.Int.2013.0239, and Terme di
517 Comano grant. CQ and NL are funded through a MRC bioinformatics fellowship (MR/M50161X/1) as part
518 of the MRC Cloud Infrastructure for Microbial Bioinformatics (CLIMB) consortium (MR/L015080/1). JTS is
519 supported by the Ontario Institute for Cancer Research through funding provided by the Government of
520 Ontario.

521

Author Contributions

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

Competing Financial Interest

The authors declare no competing financial interests.

Figure Captions

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 steps minimize fundamental sequence biases or artefacts e.g. removal of sequencing adaptors, quality trimming, removal of sequencing duplicates (using e.g. fastqc, trimmomatic [122], and Picard tools). Foreign or non-target DNA sequences are also filtered and samples are sub-sampled 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 (See **Table 4** for a detailed discussion). **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.

Figure 2. Assembly-based and assembly-free metagenome profiling. Starting from a metagenomic case-control design, we describe some of the steps needed to identify the organisms, the encoded functions and to try to links these samples' characteristics with the case/control condition. Left panel: An assembly-based pipeline, which can be fully reproduced following the commands and the code provided as a GitHub repository at https://github.com/chrisquince/metag-rev-sup is shown on the left. A read-based pipeline (right panel) using MetaPhlAn2 [88], HUMAnN2 [94], and a recent strain-level extension of the MetaPhlAn2 approach [88] is shown on the right. The raw data is available at http://metagexample.s3.climb.ac.uk/Reads.tar.gz.

Supplementary Figure 1. Example workflow for planning a metagenomics study. The advice presented here is targeted towards entry-level researchers in this area, with a particular focus on hypothesis-driven experiments, which of course may be designed very differently compared to

exploratory/hypothesis-generating studies. Key considerations for study design (blue box), sample collection (green box) and experimental procedures (yellow box) are highlighted. Understanding the potential for confounding factors, and optimization of design, can substantially improve the quality of both metagenomic sequence data, and interpretation. **Supplementary Table 1** contains further specific recommendations.

Tables and Boxes

Enrichment technique	Advantages	Limitations
Whole genome amplification [123]	• Highly sensitive - can generate sufficient DNA for sequencing from even tiny amounts of starting material. • Cost effective - can be applied directly to extracted environmental DNA, no need to isolate cells. • Non-specific and untargeted - can amplify DNA from the whole range of species present within a given sample.	• Amplification step can introduce significant biases, which skew resulting metagenomics profiles. • Chimeric molecules can be formed during amplification, which can confound the assembly step. • Non-specific – unlikely to improve proportional abundance of DNA from a species of interest.
Single-cell genomics [72]	• 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.	• Can be expensive to isolate single cells, requires specialist equipment. • Requires whole genome amplification step – see limitations above. • Biases introduced during genome amplification mean that it is usually only possible to recover partial genomes. • Prone to contamination.
Flow-sorting [124]	• High throughput means to sort cells of interest. • Targeted approach - can select specific taxa, including those that might be rare members of the microbial community.	• Expensive equipment, requiring specialist operators. • Requires intact cells. • Any cells in the sample that are attached to surfaces or fixed in structures e.g. biofilms may not be recovered. • Flow rates and sort volumes limit the number of cells that can be collected.
In situ enrichment [125]	• Simplifies microbial community structure - can make it easier to assemble genomes from metagenomics data. • Presence of particular taxa within enriched samples can give clues as to 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 favour of organisms that were able to thrive within the microcosm.
Culture/microculture [71]	• Cultured isolates can be extensively tested for phenotypic features. • Reference genomes can aid metagenomics assemblies. • Functional data can improve metagenomics annotations. • Places genomic data within its correct phylogenetic context.	• Low throughput, can be highly labor intensive. • Extremely biased - many microbes are inherently difficult to culture in the laboratory. • Unlikely to recover rarer members of a microbial community, as cultured isolate collections will be dominated by the most abundant organisms.
Sequence capture technologies [126]	• Oligonucleotide probes can be used to identify species of interest as recently demonstrated for culture-independent viral diagnostics • By focusing only on species of interest, higher sensitivity can be achieved particularly when large amounts of host contamination are present	• Capture kits can be expensive • Like PCR, capture fails when target organisms vary compared to the reference sequences used to design the probes • Genome coverage of targeted organisms can be uneven, affecting assemblies
Immunomagnetic separation [127]	• Targeted approach - can enrich specific taxa, including those that might be comparatively rare members of the microbial community • Far less expensive than many other targeted enrichment techniques such as single cell genomics or flow sorting. • Less technically challenging and time consuming than other targeted enrichment techniques.	• Requires intact cells. • Requires a specific antibody for the target cells of interest. • If target cell numbers are low, whole genome amplification may be needed following cell separation – see limitations above.
Background (e.g. human / eukaryotic) depletion techniques [128]	• Particularly useful for samples where microbial cell numbers are much lower than eukaryotic cells (e.g. biopsies) • Improves sensitivity - enhanced detection of microbial genomic data. • Lower sequence depth required to obtain good coverage of microbial genomes, reduced sequencing costs. • Relatively inexpensive, not technically challenging.	• Concomitant loss of bacterial DNA of interest can occur during processing steps, can bias subsequent microbiome profiling. • May introduce contamination.

Table 1: Summary of the advantages and limitations of methods to enrich for microbial cells/DNA before sequencing.

568

569

Dataset	Metagenomic assembly method	Assembly statistics for contigs longer than 1kb (values in parenthesis refers to perfect contigs* only)				
		# contigs	Total assembly size	Reconstruction %	N50†	% identity
Env. Mock community [55] (C.Q)	MetaSPAdes	16.22k (11.26k)	150.47M (108.39M)	80.93% (58.30%)	26.46k (25.88k)	99.86% (99.96%)
	MegaHIT	21.82k (16.67k)	146.72M (124.67M)	78.91% (67.05%)	16.94k (17.94k)	99.93% (99.98%)
HMP Mock community [2] (N.S.)	MetaSPAdes	0.72k (0.42k)	62.67M (31.95M)	95.15% (48.50%)	260.45k (178.28k)	99.98% (99.99%)
	MegaHIT	1.43k (1.14k)	62.09M (54.56M)	94.27% (82.84%)	124.02k (113.11k)	99.99% (99.99%)

570 **Table 2: Comparative evaluation of metagenomic assembly on mock microbial communities with**
571 **known composition.**

572

Sample‡	Assembler	#genes§	#matches against nr (95% identity)	# of species observed (nr at 95% identity)	Median # of single core genes	# of annotated COGs	# of annotated KEGG orthologues
Env Mock community [55] (C.Q)	MetaSPAdes	164750	154403	103	49.5	100681	91376
	MegaHIT	164146	154185	105	49	97119	91035
HMP Mock community [2] (N.S.)	MetaSPAdes	62850	61362	30	20	44625	36082
	MegaHIT	63304	61617	38	20	44289	36394
Gut sample [2]	MetaSPAdes	169399	111119	365	44.5	79414	76500
	MegaHIT	166289	109777	381	41.5	77666	75020
Ocean sample [6]	MetaSPAdes	124251	7397	118	42	51138	68633
	MegaHIT	151627	7987	110	60.5	67979	87344
Soil sample [129]	MetaSPAdes	34118	7411	86	4	10448	15312
	MegaHIT	44396	11008	132	11.5	17671	22524

* 'perfect contigs' 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.

† The N50 value corresponds to the size of the contig for which longer contigs represent at least half of the total assembly

‡ All samples have been subsampled to 50 million reads for inter sample comparability

§ total number of genes identified from the assembled contigs using Prodigal

Table 3: Comparative evaluation of metagenomic assembly of a set of metagenomes from diverse environments. Functional annotations performed as previously described [61](C.Q.,N.J.L).

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 only based upon 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	Assembly, mapping and binning are all computationally costly steps	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 only typically resolve the aggregate metabolism of the community, links with phylogeny are only possible in the context of known reference genomes
Expert manual supervision	Manual curation required for accurate binning/scaffolding, and for misassembly detection	Manual curation usually not needed, although the 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

Box 1. Limitations and opportunities in metagenomics.

Limitations of shotgun metagenomics

"Entry-level access" issues. 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, to some extent, rely 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 impact the quantitative profiling [130].

Microbial dark matter. Several members of a microbiome might have not been characterized before with culture-based methods or with metagenomics. This is regarded as microbial dark matter, and assembly-based approaches can recover part of this unseen diversity. A fraction of reads may still remain unused after assembly, and the size of this fraction is highly dependent on community structure and complexity (e.g. see the analysis reported in Table 2 and 3). It is also impacted by features such as sequencing noise, contaminant DNA, and microbes and plasmids that remain taxonomically obscure even after assembling part of their genome.

"Live or dead" dilemma. DNA persists in the environment after the death of the host cell, so the sequencing results may not be representative of the active microbial population. Compounds such as propidium monazide, which binds to 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.

Challenges in shotgun metagenomics

Integrative meta-omics. Complementing DNA sequencing with RNA, protein, and metabolomic high-throughput assays is possible with shotgun metatranscriptomics, mass-spectrometry-based metaproteomics and metabolomics [74]. Despite the potential of these technologies, 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 designs. Many shotgun metagenomic studies are cross-sectional and thus unpowered for assessing inter versus intra subject variability and microbiome temporal evolution. Tools for longitudinal settings have been developed [61] but more methods and data are needed to investigate the temporal dimension [131].

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

Supplementary Box 1. Common difficulties in Study design: problems and some recommended solutions.

Powering the study / Read depth requirements. The number of samples and sequencing depth required to be able to detect significant differences will depend on factors such as consistency of microbiome composition between different samples, the inherent microbial diversity of the samples, and effect size of the phenomenon being studied. **Solution**: These decisions can often be guided by results from previous studies in the same type of environment. In cases where this information is lacking it may be prudent to carry out preliminary marker gene-based studies to gauge the relative impact of each of the factors listed opposite.

Confounding variables and control groups. It is often very difficult to select a control group to compare against the samples of interest that is free from other confounding variables. An example of this is rodent microbiome research, where cage and animal batch effects can result in dramatic differences in microbiome composition, independent of the variable being studied [25]. Another example is the cross-sectional study of the microbiome associated with a disease for cases in which the patients cannot be sampled in the absence of active treatment. **Solution:** Current best practice is to collect as much metadata about each of the study groups as possible and factor these into the subsequent analyses when comparing groups. For clinical samples this typically includes features such as gender, age, antibiotic/medication use, location, dietary habits, and Bristol stool chart scores. For environmental samples this commonly includes associated parameters such as geographic location, season, pH, temperature etc. Further extensive advice for planning rodent microbiome studies is available [25]. Longitudinal sampling from the same patient/location can also act as an additional control, especially when longitudinal changes can be correlated with associated metadata.

Sample collection/preservation. It may be difficult to process and store all samples in exactly the same way (for example when samples are provided from a number of locations by different research groups). With longitudinal studies, samples collected at the final time point may spend less time in frozen storage prior to DNA extraction than samples collected at other time points. Such changes in sampling and preservation procedures may introduce systematic biases. **Solution:** Where possible, collection and preservation methodologies should be standardized throughout for all samples within a given study. All procedures used should also be recorded and included as pertinent metadata when carrying out subsequent data analyses. This should ideally include factors such as time

between collection and DNA extraction, length of time in frozen storage, and number of freeze-thaw cycles. For mammalian gut samples there is some evidence that storage in glycerol may result in more representative compositional results following long term frozen storage [132]. Similarly, freeze drying prior to long-term frozen storage may be a prudent approach [133].

Biomass/Contamination. Modern sequence based technologies are highly sensitive, meaning very small amounts of DNA are sufficient for sequencing. However, common laboratory kits and reagents are not sterile, meaning that any contamination that is present in these can potentially overwhelm the "real" signal in samples containing only a very low microbial biomass [34]. **Solution.** It is prudent to gauge the level of biomass present in samples before sequencing using a quantitative approach such as qPCR. Samples containing fewer than 10^5 microbial cells appear to be most impacted by background contamination [34]. Table 1 offers some approaches that may be tried in order to enrich cell numbers/DNA yields from samples prior to sequencing. Negative control samples, that have been processed using the same kits/reagents as the actual samples, should be sequenced in order to determine the types of contaminating microbes present. Sequence data derived from these contaminants might then be removed bioinformatically from the final sequence datasets. Note that the sensitivity of these negative controls can be enhanced by the use of carrier DNA [134].

Choice of DNA extraction methodology. This step can hugely impact the results of a metagenomics study. If the approach selected is not stringent enough to extract DNA from some cell types they will not be represented accurately in the subsequent sequence data. Fundamentally, the optimal type of DNA extraction approach will depend on the underlying composition of the cell types that are present within a given sample. Unfortunately this can vary greatly, even within the same type of sample (e.g. the faeces of some humans are dominated by Gram negative species with cell walls that are relatively easy to disrupt, while those of others are dominated by relatively recalcitrant Gram positive species). As a result, no one DNA extraction approach will work optimally for all sample types. **Solution**: The use of defined mock community controls [2] consisting of cultures derived from a mixture of the types of species that are common within a given environment can be a useful starting point to test the efficiency and accuracy of different DNA extraction methods. Mock communities can be optimized by including a phylogenetically diverse collection of species that are known to be commonly abundant in the sample type being studied. However, it is difficult to mimic the complexity of real microbial communities using simplified mocks, and impossible to test for the efficiency of the extraction step for unknown/uncultured organisms. Much evidence suggests that incorporating a bead-beating step into the DNA extraction process improves yield and representativeness of resulting species profiles compared to chemical-only lysis [31,135](ref#133 C.Q.,N.J.L.). However, this type of approach does typically result in more sheared DNA, potentially limiting the power of burgeoning long read sequencing technologies. DNA extraction methodology should also be included as crucial metadata when uploading sequence data to public repositories. This allows variance in methodology choices to be factored into subsequent meta-analyses that incorporate metagenomic datasets from different laboratories.

References

1 Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research* **19**, 1141-1152, doi:10.1101/gr.085464.108 (2009).

2 The 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. *Jama* **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* (2015).

9 van Kessel, M. A. *et al.* Complete nitrification by a single microorganism. *Nature*, doi:10.1038/nature16459 (2015).

10 Daims, H. *et al.* Complete nitrification by Nitrospira bacteria. *Nature*, doi:10.1038/nature16461 (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, doi:10.1016/j.cell.2014.08.032 (2014).

12 Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447-460, doi:10.1016/j.cell.2015.01.002 (2015).

13 Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe* **15**, 382-392, doi:10.1016/j.chom.2014.02.005 (2014).

14 Li, S. S. *et al.* Durable coexistence of donor and recipient strains after fecal microbiota transplantation. *Science* **352**, 586-589, doi:10.1126/science.aad8852 (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. (2012).

18 Tickle, T. L., Segata, N., Waldron, L., Weingart, U. & Huttenhower, C. Two-stage microbial community experimental design. *The ISME journal* **7**, 2330-2339, doi:10.1038/ismej.2013.139 (2013).

19 Bonder, M. J. & Kurilshikov, A. The effect of host genetics on the gut microbiome. doi:10.1038/ng.3663 (2016).

20 Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560-564, doi:10.1126/science.aad3503 (2016).

21 Knight, R. *et al.* Unlocking the potential of metagenomics through replicated experimental design. *Nature biotechnology* **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. *The ISME journal* **7**, 2116-2125, doi:10.1038/ismej.2013.106 (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 microbiology reviews*, fuv036 (2015).

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, doi:10.1038/nbt.1823 (2011).

27 Lozupone, C. A. *et al.* Meta-analyses of studies of the human microbiota. *Genome research* **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 biology* **12**, 87, doi:10.1186/s12915-014-0087-z (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. *bioRxiv*, 125724 (2017).

38 Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036, doi:10.1371/journal.pone.0128036 (2015).

39 Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 14024-14029, doi:10.1073/pnas.1519288112 (2015).

40 Simpson, J. T. & Pop, M. The Theory and Practice of Genome Sequence Assembly. *Annu Rev Genomics Hum Genet* (2015).

41 Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **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**, i94-101 (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, doi:10.1093/bioinformatics/bts174 (2012).

47 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

48 Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome research* **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. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 13272-13277 (2012).

51 Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology* **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, doi:10.1093/bioinformatics/btv033 (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 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455-477 (2012).

55 D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55, doi:10.1186/s12864-015-2194-9 (2016).

56 Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation − a benchmark of computational metagenomics software. *bioRxiv*, doi:10.1101/099127 (2017).

57 Karlin, S., Mrazek, J. & Campbell, A. M. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**, 3899-3913 (1997).

58 Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**, R85, doi:10.1186/gb-2009-10-8-r85 (2009).

59 Rosen, G., Garbarine, E., Caseiro, D., Polikar, R. & Sokhansanj, B. Metagenome fragment classification using N-mer frequency profiles. *Adv Bioinformatics* **2008**, 205969 (2008).

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

61 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144-1146 (2014).

62 Strous, M., Kraft, B., Bisdorf, R. & Tegetmeyer, H. E. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Frontiers in microbiology* **3**, 410, doi:10.3389/fmicb.2012.00410 (2012).

63 Kelley, D. R. & Salzberg, S. L. Clustering metagenomic sequences with interpolated Markov models. *BMC bioinformatics* **11**, 544, doi:10.1186/1471-2105-11-544 (2010).

64 Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome research* **23**, 111-120 (2013).

65 Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-538 (2013).

66 Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101-1106, doi:10.1126/science.aac4812 (2015).

67 Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603, doi:10.7717/peerj.603 (2014).

68 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, doi:10.7717/peerj.1165 (2015).

69 Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

70 Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 16048 (2016).

71 Stewart, E. J. Growing unculturable bacteria. *J Bacteriol* **194**, 4151-4160, doi:10.1128/jb.00345-12 (2012).

72 Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431-437 (2013).

73 Nelson, K. E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994-999 (2010).

74 Segata, N. *et al.* Computational meta'omics for microbial community studies. *Mol Syst Biol* **9**, 666 (2013).

75 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, doi:10.1038/nbt.2939 (2014).

76 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).

77 Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99-103 (2013).

78 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546 (2013).

79 Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10**, 766 (2014).

80 Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64 (2014).

81 Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome research* **21**, 1552-1560 (2011).

82 Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**, 673-676, doi:10.1038/nmeth.1358 (2009).

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

84 Xiao, L. *et al.* A catalog of the mouse gut metagenome. *Nat Biotechnol* **33**, 1103-1108, doi:10.1038/nbt.3353 (2015).

85 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).

86 Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, 1196-1199 (2013).

87 Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811-814, doi:10.1038/nmeth.2066 (2012).

88 Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902-903, doi:10.1038/nmeth.3589 (2015).

89 Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *bioRxiv*, 103085 (2017).

90 Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nature biotechnology* **33**, 1045-1052 (2015).

91 Donati, C. *et al.* Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. *Nature Microbiology*, 16070 (2016).

92 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**, e132 (2010).

93 Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, 834-841, doi:10.1038/nbt.2942 (2014).

94 Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS computational biology* **8**, e1002358, doi:10.1371/journal.pcbi.1002358 (2012).

95 Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**, D199-205 (2014).

96 UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **42**, D191-198 (2014).

97 Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212-216, doi:10.1038/nature17672 (2016).

98 Kaminski, J. *et al.* High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS computational biology* **in press** (2015).

99 Liu, B. & Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res* **37**, D443-447 (2009).

100 Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal* **9**, 207-216 (2015).

101 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

102 Oksanen, J. *et al.* The vegan package. *Community ecology package* **10**, 631-637 (2007).

103 Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**, 1200-1202, doi:10.1038/nmeth.2658 (2013).

104 Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS computational biology* **8**, e1002687 (2012).

105 Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS computational biology* **8**, e1002606 (2012).

106 Pasolli, E., Truong, D. T., Malik, F., Waldron, L. & Segata, N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS computational biology* **12**, e1004977, doi:10.1371/journal.pcbi.1004977 (2016).

107 White, J. R., Nagarajan, N. & Pop, M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology* **5**, e1000352, doi:10.1371/journal.pcbi.1000352 (2009).

108 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol* **12**, R60, doi:10.1186/gb-2011-12-6-r60 (2011).

109 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).

110 Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* **12**, 385 (2011).

111 Duy Truong, T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *in revision* (2016).

112 Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods*, doi:10.1038/nmeth.3802 (2016).

113 Quince, C. *et al.* De novo extraction of microbial strains from metagenomes reveals intra-species niche partitioning. *bioRxiv*, 073825 (2016).

114 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).

115 Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* (2015).

116 Kuleshov, V. *et al.* Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nat Biotechnol* **34**, 64-69, doi:10.1038/nbt.3416 (2016).

117 Sharon, I. *et al.* Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome research* **25**, 534-543 (2015).

118 Marx, V. Microbiology: the road to strain-level identification. *Nature methods* **13**, 401-404 (2016).

119 O'Brien, J. D. *et al.* A Bayesian approach to inferring the phylogenetic structure of communities from metagenomic data. *Genetics* **197**, 925-937, doi:10.1534/genetics.114.161299 (2014).

120 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 research* **26**, 1612-1625 (2016).

121 Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37-43, doi:10.1038/nature02340 (2004).

122 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

123 de Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. (2014).

124 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. *The ISME journal* **4**, 1352-1356 (2010).

125 Delmont, T. O. *et al.* Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Frontiers in microbiology* **6** (2015).

126 Kent, B. N. *et al.* Complete bacteriophage transfer in a bacterial endosymbiont (Wolbachia) determined by targeted genome capture. *Genome Biology and Evolution* **3**, 209-218 (2011).

127 Seth-Smith, H. M. *et al.* Generating whole bacterial genome sequences of low-abundance species from complex samples with IMS-MDA. *Nature protocols* **8**, 2404-2412 (2013).

128 Lim, Y. W. *et al.* Purifying the Impure: Sequencing Metagenomes and Metatranscriptomes from Complex Animal-associated Samples. *Journal of visualized experiments: JoVE* (2014).

129 Ofek-Lalzar, M. *et al.* Niche and host-associated functional signatures of the root surface microbiome. *Nature communications* **5**, 4950, doi:10.1038/ncomms5950 (2014).

130 Beszteri, B., Temperton, B., Frickenhaus, S. & Giovannoni, S. J. Average genome size: a potential source of bias in comparative metagenomics. *The ISME journal* **4**, 1075-1077 (2010).

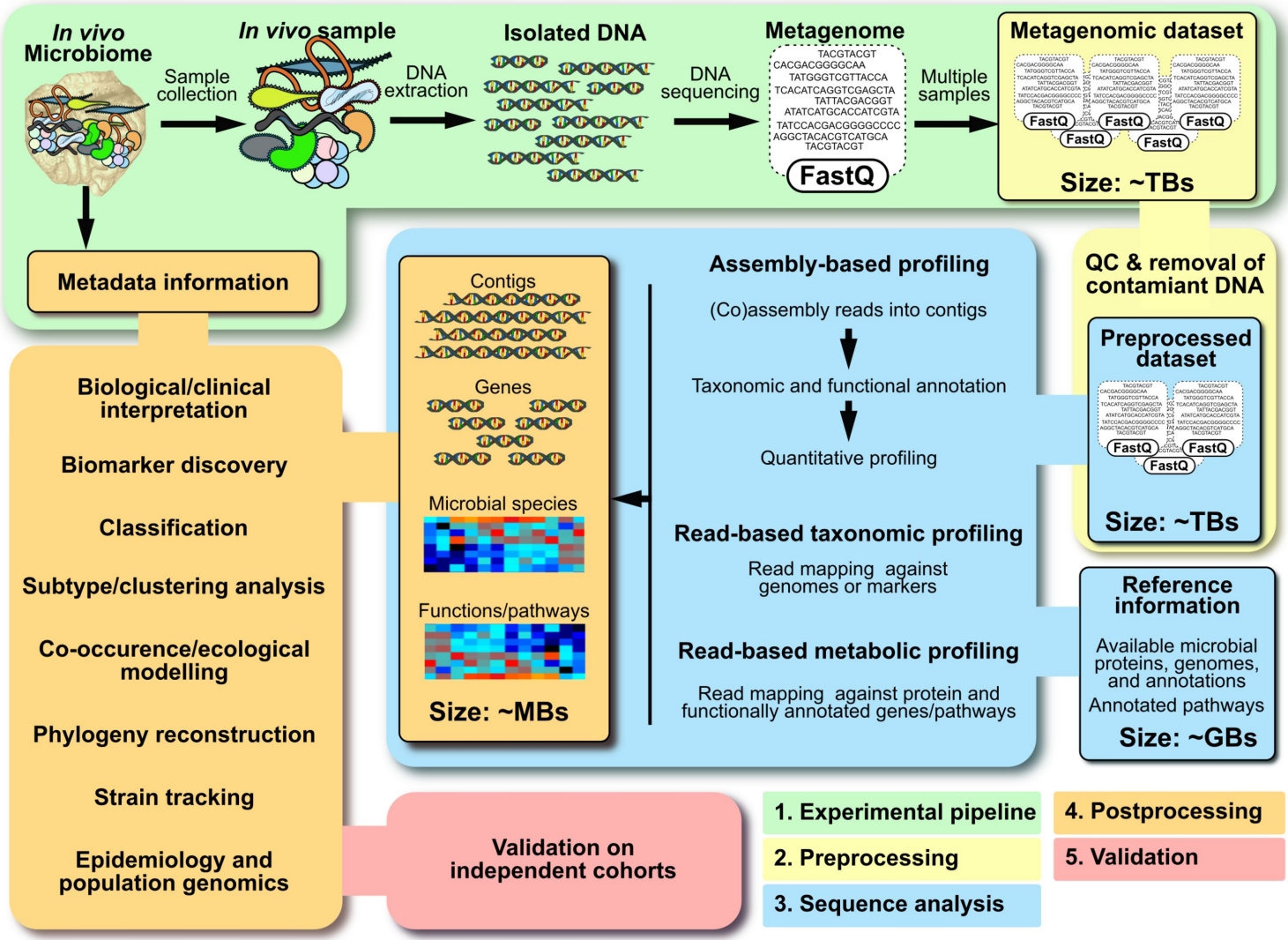
131 Knight, R. *et al.* Unlocking the potential of metagenomics through replicated experimental design. *Nat Biotechnol* **30**, 513-520, doi:10.1038/nbt.2235 (2012).

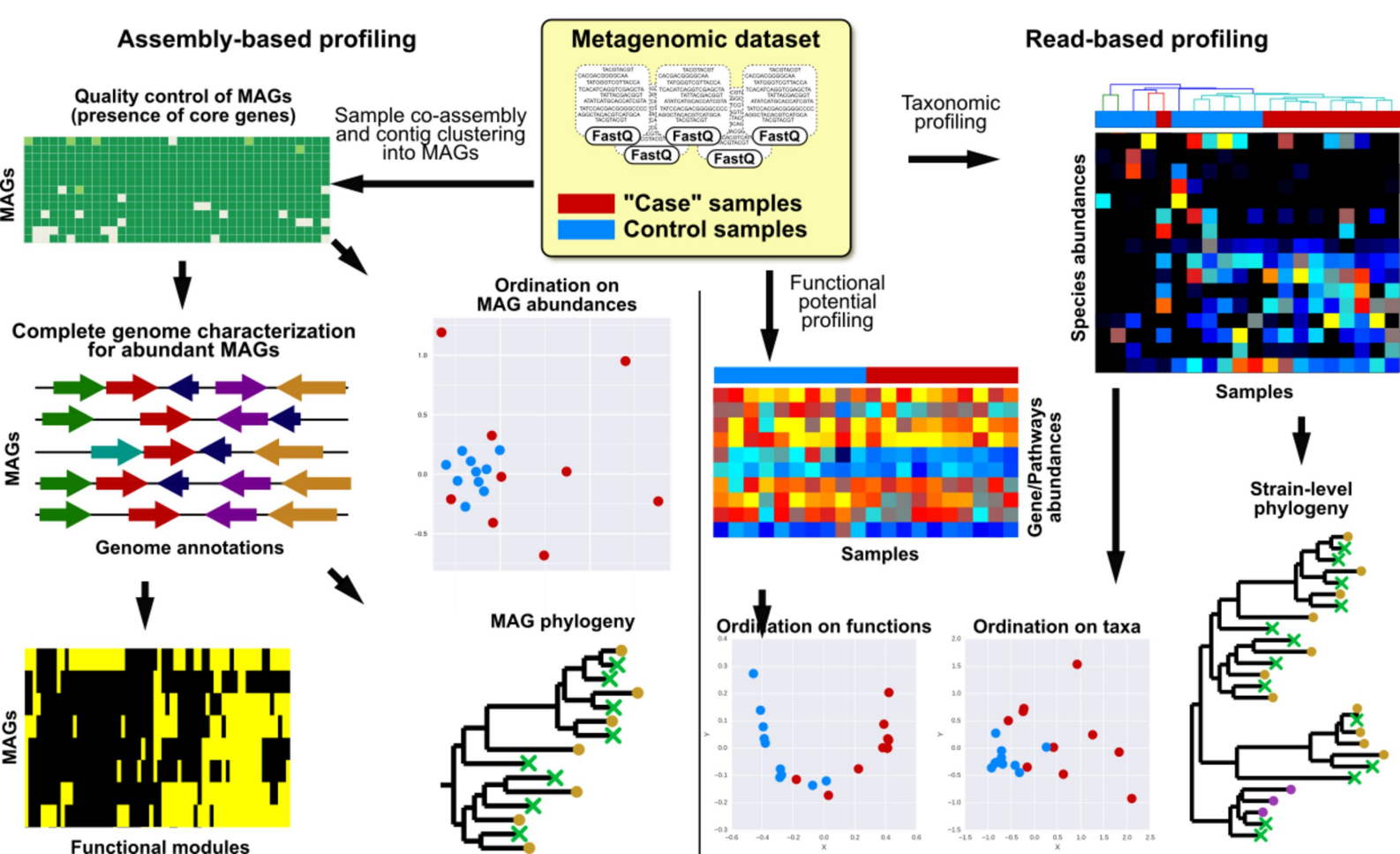
132 McKain, N., Genc, B., Snelling, T. J. & Wallace, R. J. Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. *Journal of microbiological methods* **95**, 381-383, doi:10.1016/j.mimet.2013.10.009 (2013).

133 Kia, E. *et al.* Integrity of the Human Faecal Microbiota following Long-Term Sample Storage. *PLoS One* **11**, e0163666, doi:10.1371/journal.pone.0163666 (2016).

134 Xu, Z. *et al.* Improving the sensitivity of negative controls in ancient DNA extractions. *Electrophoresis* **30**, 1282-1285, doi:10.1002/elps.200800473 (2009).

135 Gerasimidis, K. *et al.* The effect of DNA extraction methodology on gut microbiota research applications. *BMC research notes* **9**, 365, doi:10.1186/s13104-016-2171-7 (2016).


In vivo Microbiome
Sample collection
In vivo sample
DNA extraction
Isolated DNA
DNA sequencing
Metagenome
TACGTACGT
CACGACGGGGCAA
TATGGGTCGTTACCA
TCACATCAGGTCGAGCTA
TATTACGACGGGT
ATATCATGCACCATCGTA
TATCCACGACGGGGCCCC
AGGCTACACGTCATGCA
TACGTACGT
FastQ
Multiple samples
Metagenomic dataset
FastQ
Size: ~TBs
Metadata information
Contigs
Genes
Microbial species
Functions/pathways
Size: ~MBs
Assembly-based profiling
(Co)assembly reads into contigs
Taxonomic and functional annotation
Quantitative profiling
Read-based taxonomic profiling
Read mapping against genomes or markers
Read-based metabolic profiling
Read mapping against protein and functionally annotated genes/pathways
QC & removal of contamiant DNA
Preprocessed dataset
FastQ
Size: ~TBs
Reference information
Available microbial proteins, genomes, and annotations
Annotated pathways
Size: ~GBs
Biological/clinical interpretation
Biomarker discovery
Classification
Subtype/clustering analysis
Co-occurence/ecological modelling
Phylogeny reconstruction
Strain tracking
Epidemiology and population genomics
Validation on independent cohorts
1. Experimental pipeline
2. Preprocessing
3. Sequence analysis
4. Postprocessing
5. Validation


Assembly-based profiling
Metagenomic dataset
Read-based profiling
Quality control of MAGs
(presence of core genes)
MAGs
Sample co-assembly
and contig clustering
into MAGs
FastQ
FastQ
FastQ
FastQ
FastQ
"Case" samples
Control samples
Taxonomic
profiling
Species abundances
Samples
Ordination on
MAG abundances
Functional
potential
profiling
Gene/Pathways
abundances
Samples
Strain-level
phylogeny
Complete genome characterization
for abundant MAGs
MAGs
Genome annotations
MAG phylogeny
Ordination on functions
Ordination on taxa
MAGs
Functional modules