DOI:
10.1007/s00253-018-9209-9

Document Version
Publisher's PDF, also known as Version of record

Applied Microbiology and Biotechnology (2018) 102:8629–8646
https://doi.org/10.1007/s00253-018-9209-9

MINI-REVIEW

Conducting metagenomic studies in microbiology and clinical research

Tiphaine C. Martin[1,2] · Alessia Visconti[1] · Tim D. Spector[1] · Mario Falchi[1]

Received: 27 February 2018 / Revised: 28 June 2018 / Accepted: 28 June 2018 / Published online: 4 August 2018

Abstract
Owing to the increased cost-effectiveness of high-throughput technologies, the number of studies focusing on the human microbiome and its connections to human health and disease has recently surged. However, best practices in microbiology and clinical research have yet to be clearly established. Here, we present an overview of the challenges and opportunities involved in conducting a metagenomic study, with a particular focus on data processing and analytical methods.

Keywords Metagenomics · Human microbiome · Microbiology and clinical research · Next generation sequencing

Introduction

Recently, an increasing number of studies have investigated the human microbiome (bacteria, archaea, microbial eukaryotes, fungi, and viruses), particularly of the gut, and its involvement in human disease (Lynch and Pedersen 2016), including metabolic (Boulangé et al. 2016), autoimmune (Proal et al. 2009), and neuropsychiatric (Kang et al. 2013) disorders. Indeed, the human gut microbiome is involved in many host functions, such as the production of enzymes to help food digestion (Bhattacharya et al. 2015), the synthesis of vitamins (e.g., biotin – vitamin B7) and other key compounds (e.g., gamma-aminobutyric acid, Barrett et al. 2012), and the development of the host immune system (Thaiss et al. 2016). Perturbation of the gut microbiota (dysbiosis) has been associated with many diseases, as in *Clostridium difficile* infection, which is associated with a reduction of gut microbial diversity, often resulting from antibiotic use (De La Cochetière et al. 2008). The intestinal microbiota composition can also influence drug action. For instance, Dubin et al. (2016) showed that patients with metastatic melanoma having a higher proportion of *Bacteroidetes phylum* were less affected by colitis following treatment with Ipilimumab.

The manipulation of the human gut microbiome has been suggested as a potential therapeutic option for different human diseases. Human faecal microbiota transplant (FMT), which involves the transfer of faeces from a healthy donor, has been a very successful treatment for *C. difficile* infections (Eiseman et al. 1958; Gough et al. 2011; van Nood et al. 2013), and it seems to be a promising approach to treat other diseases (e.g., ulcerative colitis, Anderson et al. 2012; insulin sensitivity, Vrieze et al. 2012). Another way to modify the gut microbiome is through the administration of probiotics (suspensions of live microorganisms, AlFaleh et al. 2012) and prebiotics (substances supporting resident beneficial microorganisms, Underwood et al. (2009) and Panigrahi et al. (2017)).

Advances in DNA sequencing, triggered by the development of high-throughput sequencing technologies (or next generation sequencing—NGS), make it nowadays possible to study the diversity of microorganisms present in/on the human body in a routine and inexpensive way, and without the need for cell cultures. This allows the characterisation of microorganisms particularly hard or (so far) impossible to culture, and of previously unknown ones (Schloss and Handelsman 2005; Human Microbiome Jumpstart Reference Strains Consortium 2010; Vartoukian et al. 2010). Two main approaches are currently in use: marker

Tiphaine C. Martin and Alessia Visconti contributed equally

✉ Tim D. Spector
tim.spector@kcl.ac.uk

✉ Mario Falchi
mario.falchi@kcl.ac.uk

Tiphaine C. Martin
tiphaine.martin@kcl.ac.uk; tiphaine.martin@mssm.edu

Alessia Visconti
alessia.visconti@kcl.ac.uk

[1] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK

[2] Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY USA

gene amplification (*metagenetics*, Handelsman 2009), and whole genome shotgun sequencing (*metagenomics,* Almeida and Pop 2015). Metagenetics approaches use a polymerase chain reaction (PCR) amplification of certain phylum-specific genes, (e.g., 16S ribosomal RNA (rRNA) for bacteria and archaea and 18S rRNA for fungi), followed by their sequencing. However, uncertainty exists in the accuracy of annotations for genus and species level, especially for those organisms that have not been well characterised yet. Metagenetics approaches are cost-effective and have been widely used for studying the association between microbiome abundance and several human traits or diseases (Shreiner et al. 2015). Metagenomics approaches, by sequencing the whole genome of all the microorganisms present in a sample, are tenfold more costly but allow to potentially infer the taxonomic profiles up to the strain level, thus allowing a deeper understanding of the physiology and ecology of the microbial community. In 2008, the Human Microbiome Project (Methé et al. 2012) and the Metagenome of the Human Intestinal Tract (MetaHIT) study (Qin et al. 2010) started to characterise and generate the reference genomes of bacterial strains commonly found in association with both healthy and diseased individuals. Along with DNA sequencing, microarray biochips specific for human microbiome, also known as phylogenetic microarrays or phylochips (Walker 2016), are nowadays available for the relative quantification of known microbiota.

Neither metagenomics nor metagenetics approaches can provide information on which microbial pathways are actually active. Indeed, DNA present in a sample could come

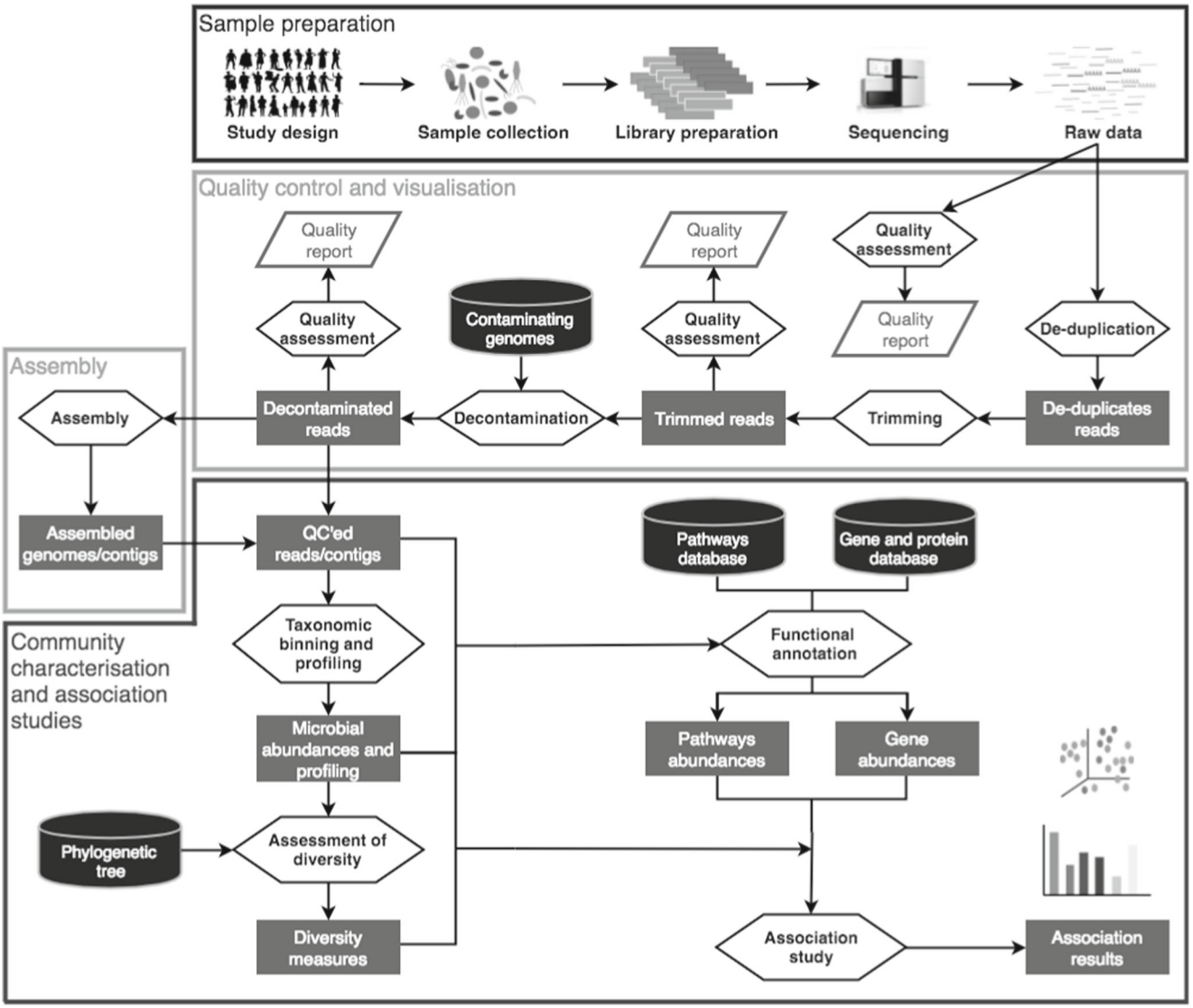

Fig. 1 Metagenomics analysis pipeline. Hexagons represent the analysis steps. Rectangles and parallelepipeds denote the output data and reports, respectively. Cylinders represent additional data to be provided in input

from resident metabolising organisms, partially quiescent cells, host cells, viruses, spores, or dead microbiota. Recently, three other high-throughput technologies have emerged: (1) *meta-transcriptomics*, that measures the expression of RNA molecules through RNA-seq (Bashiardes et al. 2016); (2) *meta-proteomics*, that measures the microbial protein levels (Grassl et al. 2016); and (3) *meta-metabolomics*, that measures the microbial metabolite levels (Zhang and Davies 2016). These technologies, combined with metagenomics, allow a better characterisation of the physiological behaviours and dynamics of the microbial community, and of their role in human health and disease.

Metagenomics is a novel and rapidly developing discipline. Therefore, standardised protocols are currently lacking, especially for the data processing and analysis, which require high computational resources and bioinformatics expertise. In this review, we will discuss best practices for the implementation of a metagenomics project, summarised in Fig. 1, with an emphasis on quality control, a critical step often poorly described in the literature.

Sample collection and storage

Samples could be collected at virtually any body site, the most studied being the gut. Collection approaches, microbiota composition, and biomass (the overall microbiota quantity) differ markedly between sites (Costello et al. 2009). While metagenetics approaches allow inference of the taxonomic profile using small amounts of DNA, metagenomics studies require higher amounts in order to get a reasonable coverage of all the present microbial genomes. For instance, in the comparative study performed by Ranjan et al. (2016), the authors used 50 ng of microbial DNA for the 16S rRNA amplicon library preparation but 5 μg for the shotgun whole metagenome one. Although several efforts have been made to improve sequencing using smaller quantities of DNA, particularly for samples with low biomass (e.g., skin metagenome), a reduced DNA quantity can affect the inferred microbiome composition (Bowers et al. 2015).

The protocols used to select, store, prepare, and sequence the samples should be consistent throughout the project to avoid the introduction of unwanted technical variability that would be difficult to remove afterwards. For instance, Sinha et al. (2016) compared different faecal sample collection methods, and concluding that all of them showed high reproducibility, although sampling methods affected the observed microbiota variability. Shaw et al. (2016) investigated the effect of sample storage and preparation, concluding that neither the duration of long-term freezing at − 80 °C nor the storage at room temperature for less than 2 days significantly affected the microbial community composition, thus suggesting that samples should be shipped on the day of collection and then processed or frozen at − 80 °C within 2 days. However, Amir et al. (2017) showed that room temperature storage is associated with a bloom of certain bacteria, which alters the taxonomic profile. Also, Choo et al. (2015) showed that while refrigeration at 4 °C do not significantly alter faecal microbiota diversity or composition, other preservative buffers (namely RNAlater, OMNIgene.GUT, and Tris-EDTA) do.

Extensive clinical and demographic data should be collected along with the specimen sample (Methé et al. 2012). Indeed, several factors, including the geographical location where the subjects live, their body mass index, and their age, have been observed to play a role in the composition of the microbial community (Yatsunenko et al. 2012; Zhernakova et al. 2016). Stool consistency (measured by the Bristol Stool Chart (Lewis and Heaton 1997) and considered a proxy for intestinal transit time) should also be recorded, since it has been associated with species richness and community composition (Tigchelaar et al. 2016; Vandeputte et al. 2016). Koren et al. (2012) also observed dramatic changes in the third trimester of pregnancy, when the gut microbiome resembles that of subjects affected by metabolic syndromes. Diet is another important factor. While long-term dietary habits are firmly associated with the microbial composition (Wu et al. 2012), it has been shown that the short-term consumption of exclusively plant- or animal-based diets (David et al. 2014) and the dietary fluctuations between seasons (Davenport et al. 2014; Smits et al. 2017) can alter the microbial community structure. When collecting infant samples, the type of delivery and whether the baby has been breast- or bottle-fed should be taken into account (Bäckhed et al. 2015), while, when collecting samples from the female reproductive tract, the age of menarche, the number of pregnancies, the menopausal status, and the type and the duration of hormonal drug intake should also be collected (Markle et al. 2013). Short-term exposure to antibiotics alters both the bacterial physiology and the microbial community structure (Maurice et al. 2013), as with many other drugs, such as metformin (Forslund et al. 2015), analgesics (Pumbwe et al. 2007), and proton pump inhibitors (Jackson et al. 2016). Finally, in the design of a case control study, cases and controls should be carefully matched for any variable that may affect the microbiome composition.

Several reviews have been written on general experimental design (e.g., Kreutz and Timmer 2009), how to choose study samples and controls (e.g., Goodrich et al. 2014), and how to select, preserve, and prepare samples before sequencing (e.g., Lauber et al. 2010; Dominianni et al. 2014; Choo et al. 2015; Voigt et al. 2015). Although, the majority of them focus on metagenetics data, their guidelines are useful for the design of metagenomics experiments.

DNA extraction and library preparation

Prior to sequencing, particular attention should be paid to the DNA extraction and library preparation. For instance, both temperature of sequencing and DNA extraction procedure can make it difficult to sequence, within the same experiment, organisms that are characterised by different GC content and/or cell membrane composition (Bohlin et al. 2010; Peabody et al. 2015; Bag et al. 2016). Indeed, microbial cells membranes are highly heterogeneous and different lysing methods can extract different amount of DNA from different species, thus generating spurious differences in their abundances when assessed from sequencing data (Bag et al. 2016). This has been confirmed in a recent analysis which compared 21 DNA extraction protocols on the same faecal samples, showing that the DNA extraction step has a large effect on the quantification of the microbial community, therefore jeopardising comparability and replicability of research findings (Costea et al. 2017).

During the library preparation step, adapter sequences (i.e., short synthetic oligo-nucleotides, often platform-specific) are ligated to the 5' and 3' ends of each DNA fragment, and often amplified. Adapters include a PCR primer binding site for amplification, and, possibly, a barcode used when multiple samples are sequenced together on the same lane. Library preparation can also affect the abundance of some DNA sequences, and, therefore, of the inferred microbiome community composition, mostly due to differential efficiency in their amplification (Aird et al. 2011). For instance, it has been observed that GC-rich DNA sequences are more difficult to amplify, and that the higher the CG content the lower the probability of an amplification bias (Jones et al. 2015). Thus, the adoption of clonal-free and PCR-optimised (Aird et al. 2011) or PCR-free (Jones et al. 2015) approaches have been recommended. Moreover, this effect has been shown to be stronger in low biomass samples. Indeed, the amount of starting material influences the overall read quality, which improves with the increase of the quantity of DNA in input (Bowers et al. 2015).

Sequencing technologies

From the advent of the Sanger platform, sequencing technologies have constantly been evolving, allowing a steady decrease in sequencing costs. Sanger sequencing (Sanger and Coulson 1975) generates long reads (> 700 bp) with a low sequencing error-rate (less than 0.1%). Its high per-base cost (more than 6,000 USD per Gb) and the complex and long sample preparation make its use difficult in routine clinical settings, and, nowadays, Sanger sequencing is mainly used to validate findings from NGS studies. NGS technologies have a much lower per-base cost than Sanger (50 − 500 USD per Gb), but a higher sequencing error rate (approximately 0.1–1% Ronaghi 2001; Morozova and Marra 2008). NGS technologies allow sequencing of one (single-) or both (paired-) ends of a DNA fragment, the latter being more precise but also slightly more expensive, and enabling the sequencing of only half of the reads at the same genomic coverage. Currently, 100 to 150 bp-long paired-reads generated using the Illumina 2500-HiSeq and 4000-HiSeq are considered the standard for metagenomics studies. However, it has been observed that short-sequence libraries (< 200 bp) may alter the phylogenetic and functional characterisation of microbial communities (Wommack et al. 2008; Carr and Borenstein 2014). This potential alteration is due to the high sequence homology among certain microorganisms, which may lead to misclassification (Koonin and Galperin 2003; Janda and Abbott 2007). To overcome this issue, one could increase the sequencing coverage (i.e., the number of times each DNA fragment, and, thus, the genome, is sequenced), or the read length. However, it should be kept in mind that, while increasing the read depth increases the number of taxa detected, it may also augment spurious assignments and artefacts (Jovel et al. 2016). Single-molecule sequencing generates longer reads (1,000–10,000 bp), which can facilitate the assembly of new genomes and the identification of novel bacterial species and strains (Koren et al. 2013; Kuleshov et al. 2016). Nanopore-based single-molecule sequencing also offers cloud-based bioinformatic analyses from anywhere and in real time, allowing the detection of specific microbiota in less than 6 h from the sample collection (Greninger et al. 2015). Therefore, data can be generated and analysed on the field, with potential for clinical diagnostics and public health—as seen in the West Africa Ebola (Quick et al. 2016) and Brazilian Zika (Quick et al. 2017) outbreaks. Single-molecule sequencing, however, has lower sequencing depth, higher costs (around 2,000 USD per Gb) and higher sequencing error rates (10–20%, Brown et al. 2017). Several authors proposed combining reads from both NGS and single-molecule sequencing to overcome each other limitations (Frank et al. 2016; Mende et al. 2016).

Controlling batch effects

Batch effects are technical sources of variation, common in large-scale high-throughput experiments, which are unrelated with the biological and scientific variable under

study. Many sources of technical variability can be easily avoided, for example by adopting homogeneous sampling and storage methods, DNA extraction or library preparation protocols. However, batch effects may be generated when samples are divided into different sequencing groups, which are often run at different dates (Leek et al. 2010). To avoid spurious associations, samples should be randomised across the batches for any of the variables under study and any potential confounder/covariate (Taub et al. 2010). For example, it is important to avoid enriching sequencing batches for female/males samples, or for older/younger subjects, and, if studying a disease, is essential to have balanced proportions of cases and controls across batches. Finally, it is preferable to avoid sequencing in multiple small batches each including few samples, and, when possible, to sequence all the batches within a limited period of time. Reproducibility of sequencing data (and of the results) can be additionally improved by adding both negative and positive controls (e.g., synthetic communities) in each sequencing batch or sample. This will help in calibrating measurements, and normalising data, allowing the comparison among samples by adjusting for individual technical variability (Leek et al. 2010; Jones et al. 2015).

Quality control

Quality control (QC) is crucial for generating high-quality data by identifying and removing low-quality biological samples and/or reads, and technical artefacts (Leek et al. 2010), and for improving the read mapping to reference databases, the quality of *de novo* assemblies, and the accuracy of the microbial diversity and abundance estimation (Bokulich et al. 2013; Zhou et al. 2014).

Visualisation of QC metric

Researchers can take advantage of several tools (even if not specifically designed for metagenomics data) to obtain a preliminary overview of reads' quality and to choose parameters to be used in the following QC steps. It is also important to examine the QC metrics after each QC step, in order to evaluate their effectiveness in generating high-quality QC'ed reads.

Close attention should also be paid to k-mers, i.e., all the possible sub-sequences of length *k* contained in the reads, as they could highlight low-complexity or repeated sequences (Plaza Onate et al. 2015). Although no specific guideline has been currently defined, if the k-mers distribution is not uniform and a set of k-mers is found in more than 1% of all reads, an in-depth investigation should be performed to understand their origin.

`FastQC` (Andrews 2010) is an excellent software to assess and visualise sequence quality.

De-duplication

Identical duplicated reads are usually considered as technical artefacts (Xu et al. 2012), since they are often the result of sequencing multiple copies of the same DNA fragment amplified during the PCR step (*artificial duplicates*, Ebbert et al. (2016)). Since in metagenomics the number of reads is used as an abundance measure, artificial duplicates may cause overestimation of the abundance of taxa, genes, and functions. On the other hand, *natural duplicates* (reads deriving from either the same region of different microbial clones, or from regions shared within/between multiple organisms, such as ortholog and paralog regions, and regions of DNA horizontal transfer) may also be present, and their removal may introduce underestimation of abundance (Niu et al. 2010). As a consequence, de-duplication should not be performed when using a PCR-free library as, in that case, all the duplicates will be natural duplicates. Also, duplicates should be removed before quality trimming, as it modifies the read sequence, potentially masking true duplicates or generating false ones.

De-duplication has only recently been included as a QC step in metagenomics studies. The first shotgun metagenomic projects (e.g., Qin et al. 2010) did not perform de-duplication and the `CD-HIT-DUP` module in `CD-HIT` (Li and Godzik 2006), used by the Human Microbiome Project, was the first method developed to manage duplicated reads without mapping on reference genomes.

Several tools are now available for removing exact duplicates without mapping on reference genomes: `FASTX-Toolkit` (Gordon and Hannon 2010) and `Fulcrum` (Burriesci et al. 2012) remove duplicates only in single-end reads, whereas `FastUniq` (Xu et al. 2012), the `CD-HIT-DUP` module in `CD-HIT` (Li and Godzik 2006), and the `clumpify` tool in the `BBTools` suite (Bushnell 2015) can deal with paired-end reads as well. When using pyrosequencing reads, the `Cdhit-454` software (Niu et al. 2010), along finding exact and nearly exact duplicates, also estimates the number of natural duplicates based on the type/origin of the sample and its complexity.

Trimming

The quality of the reads is affected by sample preparation and by the precision of the sequencing instrument, leading

to sequencing errors or low-confidence calling that can, in turn, influence the estimation of microbial diversity and taxonomy (Bokulich et al. 2013). The trimming step can identify and remove bases that have been called with a low-quality score, as measured by the PhRED quality score. We suggest keeping all the bases with a PhRED quality score > 10, representing a base call accuracy of 90% (i.e., the probability of calling a base out of ten incorrectly). Besides removing the low-quality bases, the trimming step also removes adapter sequences. Reads that become too short after trimming should be removed: in fact, short reads have a low sequence complexity (evaluated as 4^N, where N is the read length) and may map on multiple genomic regions or genomes. It is recommended to remove all the reads that are shorter than 60 bp (corresponding to a complexity of 4^{60} or less, Wommack et al. (2008)). When applied to paired-end reads, the trimming step can produce *singleton* reads, i.e., reads whose mate has been removed. It is recommended to keep these singleton reads in order to retain as much information as possible. However, some bioinformatics tools cannot deal with singleton reads.

A plethora of software is available to trim adapters and low-quality bases: `FASTX-Toolkit` (Gordon and Hannon 2010), `PRINTSEQ` (Schmieder and Edwards 2011b), `Trim Galore!` (Krueger 2012) (a wrapper to `cutadapt` (Martin 2011)), `Trimmomatic` (Bolger et al. 2014), `ngsShoRT` (Chen et al. 2014), `BBDuk` (Bushnell 2015), and `AfterQC` (Chen et al. 2017).

Decontamination

The last QC step is the identification and removal of contamination, i.e., of reads that do not belong to the studied ecosystem and that can cause misassembly of sequence contigs and/or erroneous read mapping on reference databases, thus hampering the downstream analyses.

Contaminated reads often originate from the host's genome, but they can also derive by cross-contamination during sample preparation and sequencing (e.g., from other DNA samples sequenced at the same time or from DNA present in the reagents, Salter et al. (2014)), and particularly represent a problem for samples with a low biomass (e.g., the skin). It can be challenging to identify sources of contamination without *a priori* knowledge. It is always recommended to remove contaminating reads deriving from the human genome, even when a chemical removing agent was used beforehand (Qin et al. 2010; Methé et al. 2012). It should be kept in mind that many low-complexity sequences and certain features (e.g., ribosomes) are highly conserved among species and should be eventually removed from the custom dataset in order to avoid false positive matches.

Blooming bacteria (i.e., bacteria that are fast growing at room temperature) may represent another source of contamination. However, while removing blooms will decrease the noise caused by dishomogeneous storing temperature, it may also hider the identification of actual associations (Amir et al. 2017).

Tools such as `FastQ Screen` (Wingett 2011) and `Meta-QC-Chain` (Zhou et al. 2014) can help the detection of potential contaminating genomes. The reference genomes of known or inferred contaminating organisms can then be used to create custom databases guiding the decontamination of metagenomics data. Once the contamination database is defined, tools such as `Deconseq` (Schmieder and Edwards 2011a), `ProDeGe` (Tennessen et al. 2015), `KneadData` (The Huttenhower Lab 2017), and multiple tools in the `BBTools` suite (`BBMap`, `BBwrap`, `BBsplit`, Bushnell 2015) can be used to remove contaminants.

Taxonomic binning and profiling

Short reads, conserved sequences, and the absence/incompleteness of many reference genomes are some of the factors that hamper generating the complete assembly of a metagenomic sample. To help assembly and other analyses, such as functional annotations, the reads are clustered according to their sequence similarity, and/or composition and assigned to specific taxa or operational taxonomic units (OTUs), a procedure known as taxonomic binning. The identified clusters can be then used to guide the assembly and to characterise the microbiome composition and the functional profiling.

Taxonomic binning is performed using either similarity-based or composition-based approaches (Droge et al. 2012). Similarity-based approaches (e.g., `BLAST` (Altschul et al. 1990), `MEGAN` (Huson et al. 2007), `IMG/M` (Markowitz et al. 2013), `MG-RAST` (Wilke et al. 2016)) assess the local similarity of the sequences with those in reference databases and try to assign them to taxa. This information can be retrieved from general databases, such as NCBI RefSeq (O'Leary et al. 2015), Kyoto Encyclopedia of Genes and Genomes (`KEGG`, Kanehisa et al. 2016), `eggNOG` (Huerta-Cepas et al. 2016), or by specific microbial databases, such as Gene Catalog from MetaHIT (Qin et al. 2010) and `MEDUSA` (Karlsson et al. 2014). Although similarity-based approaches generate binnings that are highly accurate and specific, they cannot bin sequences from organisms whose genomic sequence is unavailable or that are conserved among multiple close organisms. Consequently, many reads may remain unassigned. Composition-based approaches (e.g., `CD-HIT` (Li and Godzik 2006), `mOTU` (Sunagawa et al. 2013), `Kraken` (Wood and Salzberg 2014), `MetaPhlAn2` (Truong et al. 2015), and `CLARK` (Ounit et al. 2015))

evaluate the similarity of the sequences to compositional signature, such as clade-specific markers, codon usage, (i.e., the frequency of occurrence of synonymous codons in the genome) GC content, and k-mers composition, using reference databases. Since composition-based approaches do not perform intensive read alignment, they are much faster than similarity-based approaches. However, they are affected by the non-uniform representation of the various taxonomic groups in existing reference databases, and by the frequency of the compositional signatures that are usually derived only from short, specific, sequences, and not from the whole bacterial genomes. A mixed approach is used by SPHINX (Mohammed et al. 2011), which, for each read, first employs a composition-based binning algorithm to identify a subset of sequences from the reference database, and then limits the similarity-based search to this subset.

The microbiome composition, or taxonomic profile, can be computed as an absolute value, thus reporting the number of reads mapping to the detected microorganisms, or as a relative value, i.e., the relative abundance of that microorganism compared to the rest of the microbial community. Several tools have been developed to estimate microbial abundances, including GRAMMy (Xia et al. 2011), Kraken (Wood and Salzberg 2014), and ConStrains (Luo et al. 2015). An efficient approach for taxonomic profiling has been implemented in MetaPhlAn2 (Truong et al. 2015), which uses clade-specific markers to both detect the organisms present in a microbiome sample and estimate their relative abundance, allowing both binning and profiling at the same time.

Assembly

Sequence assembly is the process of aligning and merging reads with the aim of reconstructing the original genomic sequence, and it is a major step for determining the sequence of novel bacterial genomes. For instance, the HMP reconstructed the reference genome sequences for at least 900 bacteria belonging to the human microbiome (Human Microbiome Jumpstart Reference Strains Consortium 2010). Metagenomic datasets are composed of a mixture of reads belonging to multiple organisms, with different levels of taxonomic relatedness with each other and most assemblers, designed to assemble single, clonal, genome, are not able to handle these complex pan-genomic mixtures (Nielsen et al. 2014). For example, they may find it difficult to assign syntenic blocks (i.e., blocks of conserved sequence) to different organisms. Obtaining the complete assembly of a particular microbiota requires the complete coverage of its genome, which is often unfeasible. For instance, when Metsky et al. (2017) studied the Zika virus epidemic, the small amount of detected reads (sequenced using the Illumina MiSeq) hindered the complete reconstruction of the Zika genome, and the authors had to apply two targeted enrichment approaches (multiplex PCR amplification and hybrid capture) to reconstruct the genome of the virus and identify its strains. Strain-specific variants represent another challenge for assembly. In fact, the rate of genetic variations between strains could be similar to the sequencing error rate, making their assembly difficult especially with a low genomic coverage.

Assembly can be either *de novo* or reference-based (Ghurye et al. 2016). *De novo* assembly combines reads into contiguous sequences without using a reference genome. Several reference-free families of methods have been proposed. Greedy approaches (e.g., TIGR Sutton et al. 1995; phrap de la Bastide and McCombie 2007) iteratively merge reads into contigs *greedily* selecting those with maximum overlaps. Overlap-layout-consensus approaches (e.g., VICUNA Yang et al. 2012; Omega Haider et al. 2014) use the pairwise overlap between reads to build a graph, that is then traversed to merge reads into contigs that are, finally, ordered and extended. Approaches based on de Bruijn graphs (e.g., MetaVelvet Namiki et al. 2012; Afiahayati and Sakakibara 2015) split reads into overlapping k-mers, organising them in a de Bruijn graph structure (de Bruijn 1946; Compeau et al. 2011) based on their co-occurrence across reads. Analogously to overlap-layout-consensus methods, contigs are generated by traversing the generated graph. The efficiency of the *de novo* approaches decreases dramatically when the sequencing-error rate increases. Greedy approaches are the fastest and most effective approaches when there are no or few repeated elements (i.e., regions where the same genomic pattern occurs in multiple times throughout the genome), and the coverage is low, while overlap-layout consensus should be preferred when high sequencing error rate is observed. None of these methods are exempt from errors, and the resulting assembly is often extremely fragmented, also due to the incomplete coverage of the bacterial sequences. The reference-based assembly is guided by the known sequence of the organism itself; if this is not available, that of the phylogenetically closest organism may be carefully used instead (e.g., in MIRA Chevreux et al. 1999; Newbler Chaisson and Pevzner 2008; AMOS Treangen 2011). Reference-based approaches are computationally faster than *de novo* ones and can potentially map both repeated regions and those with low read coverage. However, they do not cope with new sequences and complex rearrangements (e.g., translocation and inversion) or large insertion and deletion. The *de novo* and reference-based approaches can be efficiently combined in order to improve each other results (e.g., OSLay Richter et al. 2007; E-RGA Vezzi et al. 2011). Recently, Mende et al. (2016) proposed to use

single-cell NGS sequencing for improving single microbial assembling from metagenomics data.

Functional annotation

The MetaHIT project estimated that each individual's intestinal tract hosts an average of 160 microbial species (Qin et al. 2010). However, there is no consensus yet, and the real number of species and strains harboured in the human gut may be of several thousands. Following this result, Turnbaugh and Gordon (2009) explored whether a gut *core microbiome*, i.e., a group of microbes present in every human gut, existed. While they found it hard to define such a core microbiome, they identified a common core at the gene/functional level, thus highlighting the importance of studying the modification of the functional capabilities of the microbial community rather than the microbial diversity and abundance, being the former a better marker of human health.

The functional capabilities of the microbiome community can be assessed through homology-based mapping of metagenomics sequences to databases of orthologous genes or proteins with known function (Carr and Borenstein 2014). Mapping can be achieved using the Basic Local Alignment Search Tool (`BLAST`) suite, or faster BLAST-like tools (e.g., `Bowtie2` (Langmead and Salzberg 2012) and `DIAMOND` (Buchfink et al. 2014) to query gene and protein databases, respectively). The functional annotation relies on a reference database, usually chosen among the universal protein reference (`UniRef`, Suzek et al. 2015), `KEGG` (Kanehisa et al. 2016), the Protein Family Annotations (`PFAM`, Finn et al. 2014), and the Gene Ontology (Ashburner et al. 2000) databases.

Among the several pipelines available for functional annotation, a useful approach has been implemented in the `HUMAnN2` pipeline that stratifies the community in known and unclassified organisms using `MetaPhlAn2` and the `ChocoPhlAn` pan-genome database, and combines the results with those obtained through an organism-agnostic search on the `UniRef` proteomic database. `HUMAnN2` can use both the `UniRef90` and `UniRef50` databases. While `UniRef90` is, in general, the best option, as its clusters are more likely to be iso-functional and non-redundant, `UniRef50` should be preferred when dealing with poorly characterised microbiomes. In fact, in the latter case, less stringent criteria might allow mapping a larger number of reads, although at a lower resolution. The number of reads mapping to genes and proteins is converted into coverage and abundance tables, and the `MetaCyc` database is used to assess pathway abundances.

A different but also popular approach has been implemented in `MOCAT2` (Kultima et al. 2016), which carries out a preliminary assembly of the QC'ed sequence reads into larger contigs, that are then used for gene prediction with either `Prodigal` (Hyatt et al. 2010) or `MetaGeneMark` (Zhu et al. 2010). The functional annotation is finally estimated using the `eggNOG` database and integrating information from 18 publicly available resources covering different functional properties (e.g., `KEGG`, `SEED`, and `MetaCyc` for metabolic pathways, `ARDB`, `CARD`, and `Resfams` for antibiotic resistance genes, *etc*).

Methods based on reads mapping (as `HUMAnN2`) can achieve very high sensitivity, but cannot identify new genes. On the other hand, assembly based methods (as `MOCAT2`) can identify known genes and predict new one but might under-represent genes with low coverage, as multiple reads are needed to allow their assembly. Both methods are challenged by sequence homology, either by spurious mapping of the reads to multiple genes or by generating chimeric contigs through the assembly.

Gene predictions techniques (as implemented in `Prodigal`, `MetaGeneMark`, `Orphelia` (Hoff et al. 2009), `MetaGeneAnnotator` (Noguchi et al. 2008), and `GeneMark` (Besemer and Borodovsky 1999)) can be used to detect unknown genes from contigs.

Assessing diversity

The concept of ecological diversity was originally conceived by Whittaker and Whittaker (1972) who proposed three measures to compare the diversity between and within ecological environments: the α-, β-, and γ-diversities. The α-diversity measures the mean species diversity in a given ecosystem (e.g., in a metagenomics sample), while the γ-diversity measures the overall diversity of all the ecosystems under consideration (e.g., on all metagenomics samples). The β-diversity links α- and γ-diversity and measures the difference in species content between different ecosystems (e.g., between metagenomics samples). Whittaker defined β-diversity as "the extent of differentiation of communities along habitat gradients", i.e., the ratio between γ- and α-diversity (Whittaker 1960; Whittaker et al. 2001). Several metrics have been suggested for each of these diversity measures, and the most suitable α- and β-diversity metrics in metagenomics appear to be those that take into consideration the phylogenetic tree describing the evolutionary relationship among taxa, such as the Faith's phylogenetic diversity metric for α-diversity (Faith 1992) and the weighted UniFrac for β-diversity (Chang et al. 2011; Lozupone et al. 2011). α- and β- diversities can be evaluated with `QIIME` (Caporaso et al. 2010).

The MetaHIT project also showed that the classification of gut samples based on their microbial ecosystem (*enterotypes*) is not only driven by its microbial composition, but also by its functions (Arumugam et al. 2011). Therefore, gene counts and species richness should be taken into account to characterised samples according to their gene (and, thus, function) composition, which may better elucidate the microbial role in human health and disease.

Association studies

The abundances of taxa, gene/protein families, and pathways provide information about the bacterial community structure, diversity, and functions. These variables can be used to assess the association with diseases, quantitative phenotypes, or genomics features.

A rigourous quality control is particularly important to obtain reliable results from association studies and to avoid the identification of spurious associations. Low-quality samples and outliers should be identified and removed, and technical and biological variability controlled for.

In association studies, abundances can be either used as quantitative variables or recoded as presence/absence based on suitable thresholds. Moreover, abundances can be represented on a discrete scale by using the number of reads mapping to each feature, or on a truncated continuous scale by using relative abundances normalised on the total number of reads—thus being bounded between 0 and 1.

A peculiar feature of metagenomics data is the presence of many zeroes, as many organisms and functions can be either under-sampled or present only in a small subset of individuals. Consequently, their modelling requires suitable methods. In a recent review comparing several methods for the association analysis of read-count based abundances, Jonsson et al. (2016) suggested the use of generalised linear model based on an over-dispersed Poisson or negative-binomial distribution, as that used in RNA-Seq software `EdgeR` (Robinson et al. 2010) and `DESeq2` (Love et al. 2014). However, other studies have suggested that these distributions do not adequately model zero-inflated data, because the presence of zeroes might not be due to low coverage but due to the true absence of particular organisms from a large number of samples. Thus, zero-inflated negative-binomial or hurdle models (Mullahy 1986) are likely to better reflect the distributional properties of the metagenomics data (Xu et al. 2015). The analysis of relative abundances also requires specific models, since these data are bounded between 0 and 1, which means that the effect of the explanatory variables might be non-linear and that the variance might decrease at the boundaries of the distribution. Some approaches rely on the arcsine square root transformation to stabilise the variance and normalise proportional data, such as in the `MaAsLiN` multivariate statistical framework (Morgan et al. 2012). However, the use of this transformation has repeatedly been debated (e.g., Warton and Hui 2011). An alternative approach is to model the association by using zero-inflated Beta models (Ospina and Ferrari 2012).

Studying the association of one organism at a time is a reductionist approach that ignores the interactions within the bacterial community. Machine learning methods can be applied to the bacterial community as a whole to reconstruct multi-category classifications that are then associated with the trait of interest (Statnikov et al. 2013), although they still need to be improved to consider the hierarchical nature of the data. Several multivariate testing methods taking into account the phylogeny among taxa (e.g., `PERMANOVA` McArdle and Anderson 2001; `MiRKAT` Zhao et al. 2015; `aMiSPU` Wu et al. 2016) are becoming more popular as they have higher statistical power in aggregating multiple weak associations and can reduce the burden of multiple testing correction.

Finally, multi-level functional data, such as meta-transcriptomics, meta-proteomics, or meta-metabolomics, can be integrated into genome-scale metabolic models, to disentangle the metabolism of the microbial ecosystem and its interactions with the host (Orth and Thiele 2010; Shoaie and Nielsen 2014).

Computational resources

An accurate assessment of the required computational resources is essential to efficiently and successfully tackle metagenomics projects. Metagenomics data processing requires good CPU and memory resources, and a substantial amount of disk space. To help estimate computational requirements, we provide here figures derived from a simple metagenomics analysis pipeline, which has been developed to rapidly and efficiently analyse a large number of gut metagenomic samples from the TwinsUK cohort (http://twinsuk.ac.uk/). MAP, which is available at https://github.com/alesssia/MAP, implements the QC steps listed in this review and processes the raw sequence up to the generation of the microbiome abundances. Briefly, MAP performs the QC using multiple tools from the `BBmap` suite (Bushnell 2015) to remove exact duplicates, trim low-quality bases and adapter, and to remove human decontamination. Each QC step is followed by the visualisation of the data quality metrics, carried out using FastQC (Andrews 2010). MAP takes advantage of `MetaPhlAn2` (Truong et al. 2015) for fast and efficient abundance profiling, while `QIIME` (Caporaso et al. 2010) is used to evaluate multiple diversity

Table 1 Resource usage. The reported figures were obtained by applying the proposed metagenomics pipeline to 842 raw paired-end FASTQ files with an average 26M reads per sample. Experiments were run on an HPC facility using 4 threads and limiting the available RAM to a maximum of 32 GB

Step	Data format	Tool	Virtual memory peak (average [min–max])	Time (average [min–max])	Storage (average [min–max])
Raw data	(Compressed) FASTQ	–	–	–	4.48 GB [1.45-9.32GB]
Quality assessment	html + text	`FastQC`	385.53 MB [326.00–492.90 MB]	4 min 12 s [1 min 49 s–7 min 28s]	1.05 MB [0.80–1.24 MB]
De-duplication	(Compressed) FASTQ	`Clumpify`	27.74 GB [18.10–31.40 GB]	16 min 35 s [6 min 12 s –38 min 50 s]	3.26 GB [1.05–8.18 GB]
Trimming	FASTQ	`BBduk`	8.43 GB [8.00–10.90 GB]	9 min 11 s [3 min 15 s–25 min 24 s]	13.81 GB [5.23–27.42 GB]
Decontamination	FASTQ	`BBwrap`	16.82 GB [15.50–23.90 GB]	30 min 13 s [6 min–59 min 6 s]	13.80 GB [5.23–27.40 GB]
Taxonomic binning and profiling	text + biom	`MetaPhlan2`	1.52 GB [1.30–2.50 GB]	18 min 4 s [2 min–35 min 21 s]	93.12 MB [21.77–281.74 MB]

measures. Table 1 reports figures on RAM, disk occupation, and time of execution obtained using MAP to process 842 compressed raw paired-end FASTQ files, obtained using the Illumina HiSeq 2500 platform, with average 26M reads per sample. Experiments were run on a High-Performance Computing (HPC) facility using four threads and limiting the available RAM to a maximum of 32 GB. First, it is worth noting that, while some of the implemented steps need less than 4 GB of RAM, others demand as much as 32 GB (Table 1). In fact, the higher the quantity of RAM, the more the reads that can be kept in the working memory, and the less the time consumed to use the disk as virtual memory. At the same time, multiple CPUs would allow for parallelization, further speeding up the data processing. Further parallelisation can be reached using an HPC facility, where several multi-core computers are aggregated. Attention should be paid to the disk occupation. For instance, each compressed file used for this experiment occupied 1–9 GB, while during the analysis this figure increased to 11–60 GB per sample (Table 1). At the end of the processing, deleting the temporary files released 6–35 GB of disk space per sample, and the compression of the QC'ed FASTQ files freed about 60% of extra space. The disk space requested by the files generated through this pipeline is roughly seven times the size of the original samples, and the problem would exacerbate when multiple samples are processed simultaneously, as in multi-core machines or HPC facilities.

If the user does not have the necessary computational resources or the expertise to install and run software on their local machine or cluster, they can take advantage of several user-friendly pipelines (e.g., YAMP, Visconti et al 2018; MOCAT2, Kultima et al. 2016), or of multiple web resources that are being made available for metagenomics studies (Dudhagara et al. 2015). However, the latter are less customisable, and may have constraints in their utilisation.

Data sharing

Journals and funding agencies have been increasingly requiring data sharing. Data are usually uploaded in public databases, such as NCBI, EBI, and DDBJ (http://www.insdc.org/, https://www.ncbi.nlm.nih.gov/genbank/metagenome/). An increasing number of tools (e.g., `MG-RAST`, `EBI metagenomics`) are now including the functionalities to upload data into these archives, that are then shared via the The Human Pan-Microbe Communities (HPMC) database (http://www.hpmcd.org/, Forster et al. (2016)).

While data sharing policies have increased the number of publicly available datasets, researchers should still be cautious in integrating and meta-analysing data from multiple studies, as integration might be hampered by technical and biological variabilities due to study-specific protocols for sample collection, processing, and data analysis (Lozupone et al. 2013). This highlights the necessity of standardised protocols and a first effort towards this direction is represented by the BioSharing portal that defines standards for the meta-data of a wide range of *omics* datasets (http://biosharing.org).

Clinical translational

The increased feasibility of large-scale metagenomics studies is opening new avenues for answering common

microbiology questions, including understanding what species inhabit a particular environment, what they do, and their involvement in human diseases. These findings can then be translated into the clinic. For example, probiotic administration has been shown to reduce the incidence of severe necrotising enterocolitis and mortality (AlFaleh et al. 2012) and sepsis (Panigrahi et al. 2017) in preterm babies as well as lower-respiratory tract infections in infants (Szajewska et al. 2017), while the addition of fructo-oligosaccharides (which act as prebiotics) to the supplement helps the bacterial colonisation (Underwood et al. 2009). Analogously, Osborn and Sinn (2013) suggested that a simple administration of prebiotics can prevent eczema in infants, while Hsiao et al. (2017) indicate that a combination of probiotic and peanut immunotherapy can prevent allergic immune response. More research in beneficial microbe cocktails and in prebiotics supporting their growth is thus needed to increase the spectrum of diseases that could not only be treated but also prevented.

While FMT has been pioneered in the 1950s, and it is now routinely used to treat recurrent *C. difficile* infections, it still offers challenges, as, for instance, in using metagenomics screening to make FMT perfectly safe for the transplant recipients.

As the gut microbiota contributes to the metabolism of drugs modulating drug efficacy and toxicity (Carmody and Turnbaugh 2014; Dubin et al. 2016; Wilson and Nicholson 2017; Zitvogel et al. 2018), efficient bacterial composition screening would be also useful to assess whether a patient should receive a specific course of treatment, in order to check for its effectiveness and to prevent serious side effects.

Infections are becoming harder to treat with the antibiotics currently available because of the presence of antibiotic resistant bacteria, and antibiotic resistance has been named by WHO as one of the greatest threat to public health (http://www.who.int/mediacentre/news/releases/2014/amr-report/en/). Metagenomics studies may be able to unveil the evolution of antibiotic resistance, by detecting the antibiotic resistance profile of the microbial community (Forslund et al. 2014). This may also provide a way of identifying unknown unculturable bacterial carriers of antibiotic resistant genes that may potentially be horizontally transferred to other microorganisms (Huddleston 2014).

Metagenomics approaches have also been successful in monitoring and tracking infections outbreaks (e.g., Ebola (Quick et al. 2016) and Zika (Faria et al. 2017; Quick et al. 2017) viruses). However, while these works have been proven very useful *a posteriori*, more should be done to forecast the transmission routes and prevent the disease spread. For instance, Faria et al. (2016) hypothesised that there has been a single introduction of the Zika virus into the Americas in the second half of the 2013, more than one year before its detection in Brazil. Therefore, early genetic screening could have discovered and isolated the infection months in advance, possibly avoiding or limiting its spread. It is to be hoped that the decreasing costs of sequencing coupled with a deeper understanding of the microbiome will make metagenomics testing a routine screening for infectious diseases surveillance.

Despite these promising results in specific areas of high impact, some caution is necessary. We still have an incomplete understanding of what is a healthy or dysbiotic environment, and we often cannot distinguish whether it is the disease changing the microbiome or the microbiome that is responsible for the disease. More research is therefore needed before metagenomics findings can be safely and widely translated to the clinic (Quigley 2017).

Conclusion

This review presents an overview of best practices in metagenomics, with a particular focus on the methodological and computational strategies and challenges.

While a benchmark for the assessment of computational metagenomics software has been made available as part of the Critical Assessment of Metagenomic Interpretation (CAMI) challenge (Sczyrba et al. 2017), standards for data generation, access, and retrieval are still needed to allow data sharing and the exploitation of the full potential of metagenomics data in microbiology and clinical research—as pursued in the genomic field with the creation of Global Alliance for Genomics and Healthy (http://genomicsandhealth.org).

Acknowledgements TwinsUK is funded by the Wellcome Trust and MRC. The study also receives support from the National Institute for Health Research (NIHR)- funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. AV and MF wish to acknowledge Niccolò Rossi and Richard Davies for their useful comments on the manuscript. AV would like to thank Brian Bushnell for his helpful suggestions about how to use the `BBTools` suite in a metagenomics context and for providing several useful resources.

Funding Information TwinsUK was funded by the Wellcome Trust and MRC. The study also receives support from the National Institute for Health Research (NIHR)- funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Afiahayati SK, Sakakibara Y (2015) Metavelvet-SL: an extension of the velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res 22(1):69–77

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12(2):R18

AlFaleh K, Anabrees J, Bassler D, Al-Kharfi T (2012) Cochrane review: probiotics for prevention of necrotizing enterocolitis in preterm infants. Evidence-Based Child Health: A Cochrane Review Journal 7(6):1807–1854

Almeida M, Pop M (2015) High-throughput sequencing as a tool for exploring the human microbiome. In: Metagenomics for microbiology, Elsevier, pp 55–66

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Amir A, McDonald D, Navas-molina JA, Debelius JW, Morton J, Hyde ER, RobbinsPianka A, Knight R (2017) Correcting for microbial blooms in fecal samples during room-temperature shipping. mSystems 2(2):1–5

Anderson J, Edney R, Whelan K (2012) Systematic review: faecal microbiota transplantation in the management of inflammatory bowel disease. Aliment Pharmacol Ther 36(6):503–516

Andrews S (2010) FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, Bertalan M, Borruel N, Casellas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen BH, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos MW, Brunak S, Dore J, Consortium M, Weissenbach J, Ehrlich DS, Bork P (2011) Enterotypes of the human gut microbiome. Nature 473(7346):174

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25

Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, Li Y, Xia Y, Xie H, Zhong H, Khan MT, Zhang J, Li J, Xiao L, Al-Aama J, Zhang D, Lee YS, Kotowska D, Colding C, Tremaroli V, Yin Y, Bergman S, Xu X, Madsen L, Kristiansen K, Dahlgren J, Wang J (2015) Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17(5):690–703

Bag S, Saha B, Mehta O, Anbumani D, Kumar N, Dayal M, Pant A, Kumar P, Saxena S, Allin KH, Hansen T, Arumugam M, Vestergaard H, Pedersen O, Pereira V, Abraham P, Tripathi R, Wadhwa N, Bhatnagar S, Prakash VG, Radha V, Anjana RM, Mohan V, Takeda K, Kurakawa T, Nair GB, Das B (2016) An improved method for high quality metagenomics DNA extraction from human and environmental samples. Sci Rep 6:26775

Barrett E, Ross R, O'Toole P, Fitzgerald G, Stanton C (2012) γ-Aminobutyric acid production by culturable bacteria from the human intestine. J Appl Microbiol 113(2):411–417

Bashiardes S, Zilberman-Schapira G, Elinav E (2016) Use of metatranscriptomics in microbiome research. Bioinf Biol Insights 10:19

Besemer J, Borodovsky M (1999) Heuristic approach to deriving models for gene finding. Nucleic Acids Res 27(19):3911–3920

Bhattacharya T, Ghosh TS, Mande SS (2015) Global profiling of carbohydrate active enzymes in human gut microbiome. PLoS ONE 10(11):e0142038

Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, Skjerve E, Ussery DW (2010) Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 11(1):464

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods 10(1):57

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Boulangé CL, Neves AL, Chilloux J, Nicholson JK, Dumas ME (2016) Impact of the gut microbiota on inflammation, obesity, and metabolic disease. Genome Med 8(1):42

Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T (2015) Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16:856

Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB (2017) MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience 6(3):1–10. https://doi.org/10.1093/gigascience/gix007

Buchfink B, Xie C, Huson DH (2014) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60

Burriesci MS, Lehnert EM, Pringle JR (2012) Fulcrum: condensing redundant reads from high-throughput sequencing studies. Bioinformatics 28(10):1324–1327

Bushnell B (2015) BBMap short-read aligner, and other bioinformatics tools. https://sourceforge.net/projects/bbmap/

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7(5):335

Carmody RN, Turnbaugh PJ (2014) Host-microbial interactions in the metabolism of therapeutic and diet-derived xenobiotics. J Clin Invest 124(10):4173–4181

Carr R, Borenstein E (2014) Comparative analysis of functional metagenomic annotation and the mappability of short reads. PLoS ONE 9(8):e105776

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18(2):324–330

Chang Q, Luan Y, Sun F (2011) Variance adjusted weighted UniFrac: a powerful beta diversity measure for comparing communities based on phylogeny. BMC Bioinforma 12(1):118

Chen C, Khaleel SS, Huang H, Wu CH (2014) Software for pre-processing Illumina next-generation sequencing short read sequences. Source Code Biol Med 9(1):8

Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J (2017) AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. BMC Bioinformatics 18(3):80

Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. In: German conference on bioinformatics, vol 99. Hanover, germany, pp 45–56

Choo JM, Leong LE, Rogers GB (2015) Sample storage conditions significantly influence faecal microbiome profiles. Sci Rep 5:16350

Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29(11):987–991

Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung FE, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM, Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, van Hylckama Vlieg J, Junick J, Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H, O'Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier D, Scott KP, Singh B, Slezak K, Veiga P, Versalovic J, Zhao L, Zoetendal EG, Ehrlich SD, Dore J, Bork PT (2017) Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol 35(11):1069

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R (2009) Bacterial community variation in human body habitats across space and time. Science 326(5960):1694–1697

Davenport ER, Mizrahi-Man O, Michelini K, Barreiro LB, Ober C, Gilad Y (2014) Seasonal variation in human gut microbiome composition. PLoS ONE 9(3):e90731

David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ (2014) Diet rapidly and reproducibly alters the human gut microbiome. Nature 505(7484):559–563

de Bruijn N (1946) Eenige beschouwingen over de waarde der wiskunde. Inaugural speech as professor of pure and applied mathematics and theoretical mechanics at Delft University of Technology

de la Bastide M, McCombie WR (2007) Assembling genomic DNA sequences with PHRAP. Current Protocols in Bioinformatics Chapter 11:Unit11.4

De La Cochetière MF, Durand T, Lalande V, Petit JC, Potel G, Beaugerie L (2008) Effect of antibiotic therapy on human fecal microbiota and the relation to the development of Clostridium difficile. Microb Ecol 56(3):395–402

Dominianni C, Wu J, Hayes RB, Ahn J (2014) Comparison of methods for fecal microbiome biospecimen collection. BMC Microbiol 14(1):103

Droge J, Mchardy AC, Dróge J, Mchardy AC (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. Brief Bioinform 13(6):646–655

Dubin K, Callahan MK, Ren B, Khanin R, Viale A, Ling L, No D, Gobourne A, Littmann E, Huttenhower C, Pamer EG, Wolchok JD (2016) Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. Nat Commun 7:10391

Dudhagara P, Bhavsar S, Bhagat C, Ghelani A, Bhatt S, Patel R (2015) Web resources for metagenomics studies. Genomics Proteomics Bioinformatics 13(5):296–303

Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, Duce J, Kauwe JSK, Ridge PG (2016) Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinforma 17(S7):239

Eiseman A, Silen W, Bascom G, Kauvar A (1958) Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis. Surgery 44(5):854–859

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1):1–10

Faria NR, Azevedo RdSdS, Kraemer MUG, Souza R, Cunha MS, Hill SC, Thézé J, Bonsall MB, Bowden TA, Rissanen I, Rocco IM, Nogueira JS, Maeda AY, Vasami FGdS, Macedo FLdL, Suzuki A, Rodrigues SG, Cruz ACR, Nunes BT, Medeiros DBdA, Rodrigues DSG, Nunes Queiroz AL, EVPd Silva, Henriques DF, Travassos da Rosa ES, de Oliveira CS, Martins LC, Vasconcelos HB, Casseb LMN, Simith DdB, Messina JP, Abade L, Lourenço J, Alcantara LCJ, MMd Lima, Giovanetti M, Hay SI, de Oliveira RS, Lemos PdS, LFd Oliveira, de Lima CPS, da Silva SP, JMd Vasconcelos, Franco L, Cardoso JF, Vianez-Júnior JLdSG, Mir D, Bello G, Delatorre E, Khan K, Creatore M, Coelho GE, de Oliveira WK, Tesh R, Pybus OG, Nunes MRT, Vasconcelos PFC (2016) Zika virus in the Americas: early epidemiological and genetic findings. Science 352(6283):345–349

Faria NR, Quick J, Claro I, Thézé J, de Jesus JG, Giovanetti M, Kraemer MUG, Hill SC, Black A, da Costa AC, Franco LC, Silva SP, Wu CH, Raghwani J, Cauchemez S, du Plessis L, Verotti MP, de Oliveira WK, Carmo EH, Coelho GE, Santelli ACFS, Vinhal LC, Henriques CM, Simpson JT, Loose M, Andersen KG, Grubaugh ND, Somasekar S, Chiu CY, JE Munoz-Medina, Gonzalez-Bonilla CR, Arias CF, Lewis-Ximenez LL, Baylis SA, Chieppe AO, Aguiar SF, Fernandes CA, Lemos PS, Nascimento BLS, Monteiro HAO, Siqueira IC, de Queiroz MG, de Souza TR, Bezerra JF, Lemos MR, Pereira GF, Loudal D, Moura LC, Dhalia R, França RF, Magalhães T, Marques JrET, Jaenisch T, Wallau GL, de Lima MC, Nascimento V, de Cerqueira EM, de Lima MM, Mascarenhas DL, Neto JPM, Levin AS, Tozetto-Mendoza TR, Fonseca SN, Mendes-Correa MC, Milagres FP, Segurado A, Holmes EC, Rambaut A, Bedford T, Nunes MRT, Sabino EC, Alcantara LCJ, Loman NJ, Pybus OG (2017) Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 546(7658):406–410

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) PFAM: the protein families database. Nucleic Acids Res 42(D1):D222–D230

Forslund K, Sunagawa S, Coelho LP, Bork P (2014) Metagenomic insights into the human gut resistome and the forces that shape it. Bioessays 36(3):316–329

Forslund K, Hildebrand F, Nielsen T, Falony G, Le Chatelier E, Sunagawa S, Prifti E, Vieira-Silva S, Gudmundsdottir V, Krogh Pedersen H, Arumugam M, Kristiansen K, Yvonne Voigt A, Vestergaard H, Hercog R, Igor Costea P, Roat Kultima J, Li J, Jørgensen T, Levenez F, Dore J, MetaHIT consortium, Bjørn Nielsen H, Brunak S, Raes J, Hansen T, Wang J, Dusko Ehrlich S, Bork P, Pedersen O (2015) Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature 528(7581):262

Forster SC, Browne HP, Kumar N, Hunt M, Denise H, Mitchell A, Finn RD, Lawley TD (2016) HPMCD: the database of human microbial communities from metagenomic datasets and microbial reference genomes. Nucleic Acids Res 44(D1):D604–9

Frank JA, Pan Y, Eijsink VGH, Mchardy AC (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep 6:1–10

Ghurye JS, Cepeda-Espinoza V, Pop M (2016) Metagenomic assembly: overview, challenges and applications. Yale J Biol Med 89(3):353–362

Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE (2014) Conducting a microbiome study. Cell 158(2):250–262

Gordon A, Hannon G (2010) Fastx-toolkit. http://hannonlab.cshl.edu/fastx_toolkit

Gough E, Shaikh H, Manges AR (2011) Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent clostridium difficile infection. Clin Infect Dis 53(10):994–1002

Grassl N, Kulak NA, Pichler G, Geyer PE, Jung J, Schubert S, Sinitcyn P, Cox J, Mann M (2016) Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. Genome Med 8(1):1

Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM, Dodd R, Mulembakani P, Schneider BS, Muyembe-Tamfum JJ, Stramer SL, Chiu CY (2015) Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med 7(1):99

Haider B, Ahn TH, Bushnell B, Chai J, Copeland A, Pan C (2014) Omega: an overlap-graph de novo assembler for metagenomics. Bioinformatics 30(19):2717–2722

Handelsman J (2009) Metagenetics: spending our inheritance on the future. Microb Biotechnol 2(2):138–139

Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37(suppl_2):W101–W105

Hsiao KC, Ponsonby AL, Axelrad C, Pitkin S, Tang MLK, Burks W, Donath S, Orsini F, Tey D, Robinson M, Su EL (2017) Long-term clinical and immunological effects of probiotic and peanut oral immunotherapy after treatment cessation: 4-year follow-up of a randomised, double-blind, placebo-controlled trial. The Lancet Child & Adolescent Health

Huddleston JR (2014) Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infection and Drug Resistance 7:167–176

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, Von Mering C, Bork P (2016) EGGNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44(D1):D286–D293

Human Microbiome Jumpstart Reference Strains Consortium (2010) A catalog of reference genomes from the human microbiome. Science 328(5981):994–999

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17(3):377–386

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 11(1):119

Jackson MA, Goodrich JK, Maxan ME, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD, Steves CJ (2016) Proton pump inhibitors alter the composition of the gut microbiota. Gut 65(5):749–756

Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45(9):2761–2764

Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, Fabani MM, Seguritan V, Green J, Pride DT, Yooseph S, Biggs W, Nelson KE, Venter JC (2015) Library preparation methodology can influence genomic and functional predictions in human microbiome research. Proc Natl Acad Sci 112(45):14024–14029

Jonsson V, Osterlund T, Nerman O, Kristiansson E (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics 17(1):78

Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, Perry T, Kao D, Mason AL, Madsen KL, Wong GK (2016) Characterization of the gut microbiome using 16s or shotgun metagenomics. Frontiers in microbiology 7

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG As a reference resource for gene and protein annotation. Nucleic Acids Res 44(D1):D457–D462

Kang DW, Park JG, Ilhan ZE, Wallstrom G, LaBaer J, Adams JB, Krajmalnik-Brown R (2013) Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children. PloS One 8(7):e68322

Karlsson FH, Nookaew I, Nielsen J (2014) Metagenomic data utilization and analysis (MEDUSA) and construction of a global gut microbial gene catalogue. PLoS Comput Biol 10(7):e1003706

Koonin EV, Galperin MY (2003) Evolutionary concept in genetics and genomics. In: Sequence—Evolution—Function, Springer, pp 25–49

Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Bäckhed HK, Gonzalez A, Werner JJ, Angenent LT, Knight R, Bäckhed F, Isolauri E, Salminen S, Ley RE (2012) Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 150(3):470–480

Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM (2013) Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol 14(9):R101

Kreutz C, Timmer J (2009) Systems biology: experimental design. FEBS J 276(4):923–942

Krueger F (2012) TrimGalore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M (2016) Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol 34(1):64–69

Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, Voigt AY, Zeller G, Sunagawa S, Bork P (2016) MOCAT2: A metagenomic assembly, annotation and profiling framework. Bioinformatics 32(16):2520–2523

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359

Lauber CL, Zhou N, Gordon JI, Knight R, Fierer N (2010) Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. FEMS Microbiol Lett 307(1):80–86

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11(10):733–739

Lewis S, Heaton K (1997) Stool form scale as a useful guide to intestinal transit time. Scand J Gastroenterol 32(9):920–924

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12):550

Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R (2011) Unifrac: an effective distance metric for microbial community comparison. ISME J 5(2):169–172

Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, Jansson JK, Gordon JI, Knight R (2013) Meta-analyses of studies of the human microbiota. Genome Res 23(10):1704–1714

Luo C, Knight R, Siljander H, Knip M, Xavier R, Gevers D (2015) Constrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33(10):1045–1052

Lynch SV, Pedersen O (2016) The human intestinal microbiome in health and disease. N Engl J Med 375(24):2369–2379

Markle JG, Frank DN, Mortin-Toth S, Robertson CE, Feazel LM, Rolle-Kampczyk U, von Bergen M, McCoy KD, Macpherson AJ, Danska JS (2013) Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. Science 339(6123):1084–1088

Markowitz VM, Chen IMA, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC (2013) IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Res 42(D1):D568–D573

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17(1):10

Maurice CF, Haiser HJ, Turnbaugh PJ (2013) Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell 152(1):39–50

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecol 82(1):290–297

Mende DR, Aylward FO, Eppley JM, Nielsen TN, DeLong EF (2016) Improved environmental genomes via integration of metagenomic and single-cell assemblies. Front Microbiol 7(FEB):143

Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, Gevers D, Petrosino JF, Abubucker S, Badger JH, Chinwalla AT, Earl AM, FitzGerald MG, Fulton RS, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi VR, Brooks P, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PS, Chen IMA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Durkin DunneASWMichaelJrand, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney L, Foster L, Di Francesco V, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, Kinder-Haake S, King NB, Knight R, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Liolios LiKKelvinand, Liu B, Liu Y, Lo CC, Lozupone CA, Lunsford RD, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavrommatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers YH, Ross MC, Russ C, Sanka RK, Sankar P, Sathirapongsasuti JF, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu YQ, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y, Yooseph S, Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs RA, Highlander SK, Weinstock GM, Wilson RK, Owen W (2012) A framework for human microbiome research. Nature 486(7402):215

Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, West K, Qu J, Baniecki ML, Gladden-Young A, Lin AE, Tomkins-Tinch CH, Ye SH, Park DJ, Luo CY, Barnes KG, Shah RR, Chak B, Barbosa-Lima G, Delatorre E, Vieira YR, Paul LM, Tan AL, Barcellona CM, Porcelli MC, Vasquez C, Cannons AC, Cone MR, Hogan KN, Kopp EW, Anzinger JJ, Garcia KF, Parham LA, Ramírez RMG, Montoya MCM, Rojas DP, Brown CM, Hennigan S, Sabina B, Scotland S, Gangavarapu K, Grubaugh ND, Oliveira G, Robles-Sikisaka R, Rambaut A, Gehrke L, Smole S, Halloran ME, Villar L, Mattar S, Lorenzana I, Cerbino-Neto J, Valim C, Degrave W, Bozza PT, Gnirke A, Andersen KG, Isern S, Michael SF, Bozza FA, Souza TML, Bosch I, Yozwiak NL, MacInnis BL, Sabeti PC (2017) Zika virus evolution and spread in the Americas. Nature 546(7658):411

Mohammed MH, Ghosh TS, Singh NK, Mande SS (2011) SPHINX-An algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27(1):22–30

Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol 13(9):R79

Morozova O, Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. Genomics 92(5):255–264

Mullahy J (1986) Specification and testing of some modified count data models. J Econ 33(3):341–365

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) Metavelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40(20):e155

Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto JM, Quintanilha dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, Kultima JR, Léonard P, Levenez F, Lund O, Moumen B, Le Paslier D, Pons N, Pedersen O, Prifti E, Qin J, Raes J, Sørensen S, Tap J, Tims S, Ussery DW, Yamada T, MetaHIT Consortium, Renault P, Sicheritz-Ponten T, Bork P, Wang J, Brunak S, Ehrlich SD (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol 32(8):822–828

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinf 11(1):187

Noguchi H, Taniguchi T, Itoh T (2008) Metageneannotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15(6):387–396

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745

Orth JD, Thiele I (2010) What is flux balance analysis? Nat Biotechnol 28(3):245

Osborn DA, Sinn JK (2013) Prebiotics in infants for prevention of allergy. The Cochrane Library

Ospina R, Ferrari SL (2012) A general class of zero-or-one inflated beta regression models. Comput Stat Data Anal 56(6):1609–1623

Ounit R, Wanamaker S, Close TJ, Lonardi S (2015) CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 16(1):236

Panigrahi P, Parida S, Nanda NC, Satpathy R, Pradhan L, Chandel DS, Baccaglini L, Mohapatra A, Mohapatra SS, Misra PR, Chaudhry R, Chen HH, Johnson JA, Morris JG, Paneth N, Gewolb IH (2017) A randomized synbiotic trial to prevent sepsis among infants in rural India. Nature 548(7668):407–412

Peabody MA, Van Rossum T, Lo R, Brinkman F (2015) Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinf 16(1):362

Plaza Onate F, Batto JM, Juste C, Fadlallah J, Fougeroux C, Gouas D, Pons N, Kennedy S, Levenez F, Dore J, Ehrlich SD, Gorochov G, Larsen M (2015) Quality control of microbiota metagenomics by k-mer analysis. BMC Genomics 16(1):183

Proal AD, Albert PJ, Marshall T (2009) Autoimmune disease in the era of the metagenome. Autoimmun Rev 8(8):677–681

Pumbwe L, Skilbeck CA, Wexler HM (2007) Induction of multiple antibiotic resistance in Bacteroides fragilis by benzene and benzene-derived active compounds of commonly used analgesics, antiseptics and cleaning agents. J Antimicrob Chemother 60(6):1288–1297

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, MetaHIT Consortium, Bork P, Ehrlich SD, Wang J (2010) A human gut microbial gene catalogue established by metagenomic sequencing. nature 464(7285):59–65

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JHJ, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IG, Hetzelt N, Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallasch E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Racine T, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, J Turner D, Pollakis G, Hiscox JA, Matthews DA, Shea MKO, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Wölfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keïta S, Rambaut A, Formenty P, Günther S, Carroll MW (2016) Real-time, portable genome sequencing for Ebola surveillance. Nature 530(7589):228–232

Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, Oliveira G, Robles-Sikisaka R, Rogers TF, Beutler NA, Burton DR, Lewis-Ximenez LL, de Jesus JG, Giovanetti M, Hill SC, Black A, Bedford T, Carroll MW, Nunes M, Alcantara EC, Sabino LCJ, Baylis SA, Faria NR, Loose M, Simpson JT, Pybus OG, Andersen KG, Loman NJ (2017) Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc 12(6):1261

Quigley EM (2017) Gut microbiome as a clinical tool in gastrointestinal disease management: are we there yet? Nat Rev Gastroenterol Hepatol 14(5):315–320

Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL (2016) Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. Biochem Biophys Res Commun 469(4):967–977

Richter DC, Schuster SC, Huson DH (2007) OSLAy: optimal syntenic layout of unfinished assemblies. Bioinformatics 23(13):1573–1579

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139–140

Ronaghi M (2001) Pyrosequencing sheds light on DNA sequencing. Genome 11(650):3–11

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94(3):441–448

Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol 6(8):229

Schmieder R, Edwards R (2011a) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PloS One 6(3):1–10

Schmieder R, Edwards R (2011b) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6):863–864

Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere. MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvoč M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu YW, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin HH, Liao YC, Silva GGZ, Cuevas DA, Edwards RA, Saha S, Piro VC, Renard BY, Pop M, Klenk HP, Göker M, Kyrpides NC, Woyke T, Vorholt JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC (2017) Critical assessment of metagenome interpretation – a benchmark of computational metagenomics software. Nature Methods

Shaw AG, Sim K, Powell E, Cornwell E, Cramer T, McClure ZE, Li MS, Kroll JS (2016) Latitude in sample handling and storage for infant faecal microbiota studies: the elephant in the room? Microbiome 4(1):40

Shoaie S, Nielsen J (2014) Elucidating the interactions between the human gut microbiota and its host through metabolic modeling. Front Genet 5(APR):1–10

Shreiner AB, Kao JY, Young VB (2015) The gut microbiome in health and in disease. Curr Opin Gastroenterol 31(1):69

Sinha R, Chen J, Amir A, Vogtmann E, Shi J, Inman KS, Flores R, Sampson J, Knight R, Chia N (2016) Collecting fecal samples for microbiome analyses in epidemiology studies. Cancer Epidemiology and Prevention Biomarkers 25(2):407–416

Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, Knight R, Manjurano A, Changalucha J, Elias JE, Dominguez-Bello MG, Sonnenburg JL (2017) Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 357(6353):802–806

Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, Pei Z, Blaser MJ, Aliferis CF, Alekseyenko AV (2013) A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 1(1):11

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10(12):1196

Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR Assembler: a new tool for assembling large shotgun sequencing projects. Genome Science and Technology 1(1):9–19

Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH (2015) Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932

Szajewska H, Ruszczyński M, Szymański H, Sadowska-Krawczenko I, Piwowarczyk A, Rasmussen PB, Kristensen MB, West CE, Hernell O (2017) Effects of infant formula supplemented with prebiotics compared with synbiotics on growth up to the age of 12 mo: a randomized controlled trial. Pediatr Res 81(5):752

Taub MA, Corrada Bravo H, Irizarry RA (2010) Overcoming bias and systematic errors in next generation sequencing data. Genome Med 2(12):87

Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, Dangl JL, Ivanova N, Woyke T, Kyrpides N, Pati A (2015) Prodege: a computational protocol for fully automated decontamination of genomes. ISME J 10(1):1–4

Thaiss CA, Zmora N, Levy M, Elinav E (2016) The microbiome and innate immunity. Nature 535(7610):65–74

The Huttenhower Lab (2017) KneadData. https://bitbucket.org/biobakery/kneaddata/wiki/Home

Tigchelaar EF, Bonder MJ, Jankipersadsing SA, Fu J, Wijmenga C, Zhernakova A (2016) Gut microbiota composition associated with stool consistency. Gut 65(3):540–542

Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M (2011) Next generation sequence assembly with AMOS. Current Protocols in Bioinformatics Chapter 11(SUPP.33):1–18

Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N (2015) Metaphlan2 for enhanced metagenomic taxonomic profiling. Nat Methods 12(10):902–903

Turnbaugh PJ, Gordon JI (2009) The core gut microbiome, energy balance and obesity. J Physiol 587(17):4153–4158

Underwood MA, Salzman NH, Bennett SH, Barman M, Mills D, Marcobal A, Tancredi DJ, Bevins CL, Sherman MP (2009) A randomized placebo-controlled comparison of two prebiotic/probiotic combinations in preterm infants: impact on weight gain, intestinal microbiota, and fecal short chain fatty acids. J Pediatr Gastroenterol Nutr 48(2):216

Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J (2016) Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. Gut 65(1):57–62

van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, Visser CE, Kuijper EJ, Bartelsman JF, Tijssen JG, Speelman P, Dijkgraaf MG, Keller JJ (2013) Duodenal infusion of donor feces for recurrent clostridium difficile. N Engl J Med 368(5):407–415

Vartoukian SR, Palmer RM, Wade WG (2010) Strategies for culture of 'unculturable' bacteria. FEMS Microbiol Lett 309(1):no–no

Vezzi F, Cattonaro F, Policriti A (2011) e-RGA: enhanced reference guided assembly of complex genomes. EMBnetjournal 17(1):46–54

Visconti A, Martin TC, Falchi M (2018) YAMP: a containerised workflow enabling reproducibility in metagenomics research, GigaScience. https://doi.org/10.1093/gigascience/giy072

Voigt AY, Costea PI, Kultima JR, Li SS, Zeller G, Sunagawa S, Bork P (2015) Temporal and technical variability of human gut metagenomes. Genome Biol 16(1):73

Vrieze A, Van Nood E, Holleman F, Salojärvi J, Kootte RS, Bartelsman JF, Dallinga–Thie GM, Ackermans MT, Serlie MJ, Oozeer R, Derrien M, Druesne A, Van Hylckama Vlieg JE, Bloks VW, Groen AK, Heilig HG, Zoetendal EG, Stroes ES, de Vos WM, Hoekstra JB, Nieuwdorp M (2012) Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. Gastroenterology 143(4):913–916

Walker AW (2016) Studying the human microbiota. In: Microbiota of the human body, Springer, pp 5–32

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecol 92(1):3–10

Whittaker ARH, Whittaker RH (1972) Evolution and measurement of species diversity. TAXON 21(2):213–251

Whittaker RH (1960) Vegetation of the Siskiyou mountains, Oregon and California. Ecol Monogr 30(3):279–338

Whittaker RJ, Willis KJ, Field R (2001) Scale and species richness: towards a general, hierarchical theory of species diversity. J Biogeogr 28(4):453–470

Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, Paczian T, Trimble WL, Bagchi S, Grama A, Chaterji S, Meyer F (2016) The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res 44(D1):D590–D594

Wilson ID, Nicholson JK (2017) Gut microbiome interactions with drug metabolism, efficacy, and toxicity. Transl Res 179:204–222

Wingett S (2011) FastQ screen. http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. Appl Environ Microbiol 74(5):1453–63

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15(3):R46

Wu C, Chen J, Kim J, Pan W (2016) An adaptive association test for microbiome data. Genome Med 8(56):1–12

Wu GD, Chen J, Hoffmann C, Bittinger K, Yy Chen, Sue A, Bewtra M, Knights D, Wa Walters, Knight R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H (2012) Linking Long-Term dietary patterns with gut microbial enterotypes. Science 334(6052):105–108

Xia L, Cram J, Chen T, Fuhrman J, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. PLoS One 6(12):e27992

Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S (2012) Fastuniq: a fast de novo duplicates removal tool for paired short reads. PLoS ONE 7(12):1–6

Xu L, Paterson AD, Turpin W, Xu W (2015) Assessment and selection of competing models for zero-inflated microbiome data. PLoS ONE 10(7):e0129606

Yang X, Charlebois P, Gnerre S, Coole MG, Lennon NJ, Levin JZ, Qu J, Ryan EM, Zody MC, Henn MR (2012) De novo assembly of highly diverse viral populations. BMC Genomics 13(1):475

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, Heath AC, Warner B, Reeder J, Kuczynski J, Caporaso JG, Lozupone CA, Lauber C, Clemente JC, Knights D, Knight R, Gordon JI (2012) Human gut microbiome viewed across age and geography. Nature 486(7402):222–227

Zhang LS, Davies SS (2016) Microbial metabolism of dietary components to bioactive metabolites: opportunities for new therapeutic interventions. Genome Med 8(1):1

Zhao N, Chen J, Carroll IM, Ringel-kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC (2015) Testing in microbiome-profiling studies with mirkat, the microbiome regression-based kernel association test. Am J Hum Genet 96(5):797–807

Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, Mujagic Z, Vila AV, Falony G, Vieira-Silva S, Wang J, Imhann F, Brandsma E, Jankipersadsing SA, Joossens M, Cenit1 MC, Deelen P, Swertz MA, Lifelines cohort studyand Weersma RK, Feskens EJM, Netea1 MG, Gevers D, Jonkers D, Franke L, Aulchenko YS, Huttenhower C, Raes J, Hofker MH, Xavier RJ, Wijmenga C, Fu J (2016) Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352(6285):565–569

Zhou Q, Su X, Jing G, Ning K (2014) Meta-QC-chain: comprehensive and fast quality control method for metagenomic data. Genomics Proteomics Bioinformatics 12(1):52–56

Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38(12):e132–e132

Zitvogel L, Ma Y, Raoult D, Kroemer G, Gajewski TF (2018) The microbiome in cancer immunotherapy: diagnostic tools and therapeutic strategies. Science 359(6382):1366–1370