Journal of Microbiological Methods 95 (2013) 401–414

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

High throughput sequencing methods and analysis for microbiome research

Julia M. Di Bella [a,1], Yige Bao [a,b,c,d,1], Gregory B. Gloor [b,e], Jeremy P. Burton [a,b,f], Gregor Reid [a,b,f,*]

[a] Department of Microbiology and Immunology, The University of Western Ontario, London, ON, Canada
[b] Canadian R&D Centre for Probiotics, Lawson Health Research Institute, London, ON, Canada
[c] Department of Urology, West China Hospital, Sichuan University, Chengdu, China
[d] West China School of Clinical Medicine, Sichuan University, Chengdu, China
[e] Department of Biochemistry, The University of Western Ontario, London, ON, Canada
[f] Department of Surgery (Urology), The University of Western Ontario, London, ON, Canada

ARTICLE INFO

Article history:
Received 27 June 2013
Received in revised form 13 August 2013
Accepted 13 August 2013
Available online 9 September 2013

Keywords:
Bioinformatics
Metagenome
Metatranscriptome
Microbiome
High-throughput sequencing

ABSTRACT

High-throughput sequencing technology is rapidly improving in quality, speed and cost. It is therefore becoming more widely used to study whole communities of prokaryotes in many niches. This review discusses these techniques, including nucleic acid extraction from different environments, sample preparation and high-throughput sequencing platforms. We also discuss commonly used and recently developed bioinformatic tools applied to microbiomes, including analyzing amplicon sequences, metagenome shotgun sequences and metatranscriptome sequences. This field is relatively new and rapidly evolving, thus we hope that this review will provide a baseline for understanding these methods of microbiome analyses. Additionally, we seek to stimulate others to solve the many problems that still exist with the sensitivity, specificity and interpretation of high throughput microbiome sequence analysis.

1. Introduction

Our world is dominated by prokaryotes. The total number of microbial cells on Earth is estimated to be 10^{30} (Turnbaugh and Gordon, 2008) and in the human body alone, there are up to 100 trillion organisms, which approximately equates to ten times the number of our own human cells (Savage, 1977). There are literally millions of prokaryotic species, though most have not yet been cultivated (Amann et al., 1995). It is likely that there are numerous enzymes and metabolic capabilities not previously described but encoded by the genes of these species. In the human body, bacteria play important roles in modulation of the digestive, endocrine and immune functions. With the advent of more recent culture-independent sequencing based methods, the composition and diversity of the human microbiome is being uncovered.

The earliest direct cloning of environmental microbial DNA was proposed by Lane et al. (1985), while the term 'metagenome' was proposed by Handelsman et al. (1998) to describe "the genomes of the total microbiota found in nature"; that is the whole collection of genomic information of all microorganisms in a given environment. With the advancement of technologies such as sequence- and function-based gene screening, high-throughput sequencing and metatranscriptomics, incredible insight has been achieved in studying microbiomes, including those associated with human health and disease (Hess et al., 2011; Qin et al., 2010). In this review, we aim to describe methods for metagenomic and metatranscriptomic studies in the context of the microbiome, and discuss progress and future steps in the field.

2. Considerations for study design

Three major types of experimental approaches will be discussed in this review. In amplicon sequencing, a particular gene, gene fragment or sequence is amplified and the sequence determined. This is usually done in very highly conserved genes, such as segments of the 16S rRNA gene, in order to determine which organisms are in a sample and how organisms differ with the environment. In metagenome sequencing, the entire DNA in a sample is sequenced to determine which genes are present in the sample and if the overall functionalities or pathways differ between environments. Lastly, in metatranscriptome sequencing (RNA-seq), all RNA in a sample is sequenced and proportionally analyzed to determine which transcripts are present and if genes are differentially expressed within the environment. All approaches rely on high-throughput sequencing technology, but differ in their sample preparation, sequencing approach and bioinformatic analyses.

Studies of the microbiome require high-quality data to generate valid results. This requires having a sufficient number of samples, taking samples from biologically relevant sites, controlling for confounding factors and using appropriate analytical tools. While this review will focus on the last factor, namely the tools that are available and appropriate for studying different aspects of the microbiome using high-

* Corresponding author at: F3-106, Lawson Health Research Institute, 268 Grosvenor Street, London, Ontario, N6A 4V2, Canada. Tel.: +1 519 646 6100x65256.
E-mail address: gregor@uwo.ca (G. Reid).

[1] These authors contributed equally to this work.

0167-7012/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.mimet.2013.08.011

throughput sequencing technology, the importance of collecting sufficient and appropriate samples cannot be discounted. Furthermore, the sequencing depth (how many reads can be produced during sequencing) is very important in order to resolve rare species or to detect subtle differences between samples. Each sequencing technology has a maximum capacity (described in Section 4) and the number of sequencing reads they can undertake influences how many samples can be analyzed at once, as well as the robustness of the conclusions one can draw from the data. For example, high depth of coverage allows the detection of more operational taxonomic units (OTUs) for amplicon sequencing-based experiments, the detection of more genes for shotgun metagenomic analyses and the detection of more areas of low transcription for transcriptomics. The depth of coverage that is necessary for an experiment depends on many factors including the goal of the experiment (for example, is the researcher trying to identify the most important bacteria in a sample or creating a more comprehensive view of its composition) and the diversity of the microbiome in the sample. A more detailed discussion of the role of sequencing depth can be found in Hamady and Knight (2009).

One very important factor for sample collection that is increasingly emphasized is the acquisition and recording of metadata, that is, data about the data. Metadata can include where the sample came from, the qualities of that environment, how that sample was processed and sequenced. The metadata that should be collected depends on the nature of the sample: for instance with regards to a soil sample it would be useful to record the latitude, longitude and depth where it was taken, whereas, for a human oral swab it would be useful to record age, gender and all aspects of the health or disease status. Metadata recording is increasingly important as more metagenomic and metatranscriptomic sequences are submitted to databases; it is helpful to be able to search by a particular disease state, habitat or other characteristics. It is also useful to have information stored about the methods used to obtain the sequences, such as the sequencing platform used. The Genome Standards Consortium has proposed checklists for metadata such as the Minimum Information about any Sequence (MIxS), the Minimum Information about a Marker Sequence (MIMARKS), and the Minimum Information about a Metagenome Sequence (MIMS) checklists in order to standardize the metadata collected about samples, as well as “environmental packages” for specific sample types (Yilmaz et al., 2011). Beyond submission to databases, metadata are also useful for any study where samples are being compared—the more metadata that are available, the more information a researcher has about how the samples differ and how these might affect microbiome findings.

3. Nucleic acid extraction and library preparation

3.1. DNA extraction techniques

Microbial genomic DNA extraction and purification serves as the first step of library preparation. Since most sequencing protocols require between nanograms and micrograms of DNA, efficient DNA isolation and purification is critical for downstream sequencing.

Cell lysis and subsequent DNA extraction from certain microbes, especially those living in extreme environments, can be difficult, as these organisms have rigid cell wall structures and also release stable nucleases upon cell lysis. However, since such microbes are an interesting source for bioprospecting new enzymes and other components, efforts are constantly made to perform those challenging DNA extractions. With the application of mechanical lysis (bead beating and sonicating), chemical lysis (SDS, phenol) and enzymatic lysis (proteinase K, mutanolysin *etc.*), high-quality DNA has been retrieved from a variety of environments, including soil (Hardeman and Sjoling, 2007; Pathak et al., 2009; Voget et al., 2006; Waschkowitz et al., 2009), marine picoplankton (Stein et al., 1996), saline soil (Purohit and Singh, 2009), contaminated subsurface sediments (Abulencia et al., 2006), groundwater (Uchiyama et al., 2005), hot springs and mud holes in solfataric fields (Rhee et al., 2005), Urania hypersaline basins (Ferrer et al., 2009), surface water from rivers (Wu and Sun, 2009), glacier ice (Simon et al., 2009), glacier soil (Yuhong et al., 2009), Antarctic and Arctic desert soil (Cieslinski et al., 2009; Heath et al., 2009; Jeon et al., 2009), and buffalo rumens (Duan et al., 2009).

In order to avoid wasting expensive reagents for library preparation and sequencing, the quantity and quality of the obtained DNA need to be confirmed before amplification and sequencing. Examples of quality assessment devices include the 2100 Bioanalyzer (from Agilent Technologies), fluorometers such as the Qubit (from Life Technologies), and quantitative PCR for DNA damage identification (Ginzinger et al., 2000). After verification of quality and quantity, the DNA templates can go through further preparation steps.

3.2. DNA preparation techniques for metagenomics

Although different sequencing techniques have various protocols, the initial DNA sample preparation generally includes three similar steps. First, the DNA molecules are fragmented to produce pieces that are small enough to sequence. Second, the fragments are given blunt ends to aid further processing. Last, adaptors are ligated to the fragments.

Fragmentation techniques can be mechanical or enzymatic, and the former can be further classified into nebulization, hydrodynamic shearing and ultrasonication. Nebulization is less expensive, but suffers from large losses of input material, broad ranges of fragment sizes, a risk of cross-contamination, and an inability to do parallel processing. Hydrodynamic shearing (*e.g.* HydroShear™ by GeneMachines™) uses a similar mechanism as nebulization, but uses a syringe pump to shear DNA by forcing it through small pores, in contrast to nebulization's high pressure gas. Ultrasonication instruments allow parallel sample processing and have low hands-on time and sample loss; however, they are more expensive.

Mechanical fragmentation techniques yield more random and readily controlled overlapping DNA fragments as compared to enzymatic methods (Knierim et al., 2011). However, mechanically fragmented DNA strands need repair and end polishing before adaptors can be ligated. In contrast, enzymatic fragmentation, a more recently developed method, can retrieve random fragments of the desired length with less DNA input (Loman et al., 2012a). For example, the “dsDNA Fragmentase” from New England Biolabs (NEB) creates random double-stranded breaks in DNA with its mutant *Vibrio vulnificus* nuclease and mutant T7 endonuclease, and acts at a constant rate, allowing the size of the fragments to be controlled in a time-dependent manner (Knierim et al., 2011).

After fragmentation, adaptor sequences are ligated on the ends of the fragments, allowing them to be attached to a solid surface for sequencing (for example, a tagged glass slide or bead). These adaptors are specific to the sequencing platform and allow amplification by acting as primers. A novel approach, “tagmentation” (Nextera™) employs *in vitro* transposition with a specialized “Transposome™” enzyme that fragments DNA strands, repairs ends and attaches sequence tags in the same step. The adaptors added can be platform-specific, making it an appropriate method for preparing samples for both Roche 454 and Illumina platforms (Caruccio, 2011).

3.3. RNA extraction techniques

For RNA extraction, one must consider the ubiquitous presence of ribonuclease enzymes in cells and tissues, which are released during cell lysis and can rapidly degrade RNA (Peirson and Butler, 2007). It is, therefore, important to inactivate endogenous RNases in samples and maintain a working environment free of RNase. In order to rapidly inactivate endogenous RNases, the method widely used is guanidinium thiocyanate-phenol-chloroform extraction, which simultaneously lyses the cell and inactivates RNases. RNA separation and isolation can then be achieved with kits from several companies, including Qiagen's silica column, Ambion's magnetic beads, and Invitrogen's Trizol extraction method. Since mRNA typically constitutes a very small proportion of the total RNA, ribosomal RNA (rRNA) depletion can also be performed to retrieve an mRNA library with the kits such as MICROBExpress (Ambion), mRNA-ONLY (Epicentre), Ribo-Zero (Epicentre), Ovation Prokaryotic RNA-seq System (NuGEN) (Giannoukos et al., 2012) and specific depletion of cDNA representing rRNAs (Yi et al., 2011). Reverse transcription is then performed to convert RNA into cDNA.

4. High throughput sequencing methodologies

The first-generation DNA sequencing technology was developed by Sanger et al. (1977) based on the selective incorporation of chain-terminating dideoxynulceotides and the first automatic sequencing machine (AB370) was produced by Applied Biosystems in 1987 (Liu et al., 2012). The Sanger sequencing technique completed the first bacterial genome sequencing in 1995 (Fleischmann et al., 1995), and constituted the main part of the Human Genome Project in 2001 (Collins et al., 2003), which in turn promoted further development of sequencing technology.

With the launch of the Genome Sequencer 20 system by 454 Life Sciences in 2005, second-generation sequencing techniques came into recognition with massively parallel analysis, high throughput and reduced cost. This significant advance greatly reduced the difficulty of sequencing (Metzker, 2005) and made it possible to analyze bacterial genomes in hours or days rather than months or years (Loman et al., 2012b). While most second-generation sequencing techniques rely on a sequencing-by-synthesis design, the newly emerged third-generation sequencing techniques are performed on a single-molecule basis with no initial DNA amplification step. In this section we will discuss the major systems of second- and third-generation sequencing. The features of each next generation sequencing device are summarized in Table 1.

4.1. General considerations for high throughput sequencing

One of the first questions a scientist considering using high-throughput sequencing might ask is: does this task actually require high-throughput sequencing? While the method is increasingly popular, fast, reliable and cheap, for some experiments, easier and even less expensive methods such as microarrays may be appropriate. However, the depth and relatively unbiased method of high-throughput sequencing is, in many cases, replacing microarray technologies.

The second question is which platform is appropriate for the experiment? There are many different platforms on the market, each with strengths and weaknesses. Some may generate the largest total throughput per run, while others may have better speed, read length or accuracy. The appropriate selection depends on the aims of the experiment. In some cases, a combination of techniques could provide more complete coverage of the genome.

In order to achieve efficiencies in time and cost, sequencing of microbial samples usually employs barcoding and multiplexing of samples. Understanding the complexity of your samples is critical to determine whether multiplexing is needed and how many samples can be assayed in a single run. If using a multiplexing approach, the concentration of each sample should be similar so that equal amounts of data are achieved for each multiplexed sample.

4.2. The Roche 454 sequencing technique

454 Life Sciences developed the GS 20 system in 2005, which introduced the sequencing-by-synthesis approach to DNA sequencing. In 2007, the company released the GS FLX system and its 2008 upgrade, GS FLX Titanium, produced a fivefold increase in throughput.

The Roche 454 platform (Petrosino et al., 2009; Ronaghi, 2001) employs pyrosequencing (Nowrousian, 2010). In this method, DNA libraries are fragmented to a size between 400 and 800 base pairs, ligated to adaptors, and denatured into single strands. The single strands are captured by amplification beads and amplified by emulsion PCR (Berka et al., 2010). The beads are then transferred to a picotiter plate, where each dNTP is washed over the plate one at a time, and the release of pyrophospate when one is incorporated drives a reaction, turning luciferin into oxyluciferin and generating visible light (Froehlich et al., 2010). The light signal recorded when the base is washed over the plate allows the system to determine the exact base being added. The Roche 454 system has a long read length and relatively high speed. The newest GS FLX Titanium XL+ sequencer can reach a length up to 1000 bp, with a throughput of 700 Mb. One major shortcoming is the cost per base on this platform is about 10 times more expensive than Illumina's HiSeq 2000. It also has quite low throughput and automation and relatively high error rates (Bennett, 2004; Gilles et al., 2011; Huse et al., 2007; Luo et al., 2012; Quince et al., 2009).

The use of 454 sequencing is especially problematic in homopolymeric tracts, that is, regions where one nucleotide is repeated many times, because of the way it detects incorporation of nucleotides. In homopolymeric tracts, many bases are incorporated at once and the intensity corresponding signal is in theory proportional to the number of nucleotides added. However, it is not very precise, so the number of nucleotides detected is not always the number added, resulting in insertion and deletion (indel) mutations at homopolymeric tracts. This quality makes it a less preferable option for 16S rDNA amplicon sequencing (Kunin et al., 2010). However, this technique's long read length is useful for sequencing repetitive or palindromic sequences, as well as scaffolding for metagenomics, since long read lengths are easier to assemble.

Roche 454 sequencing has been widely applied to the identification of microbiota within the human body, including the gut (Wylie et al.,

Table 1
Comparison of next-generation sequencing techniques.

Sequencing devices	Chemistry	Read length (bp)	Run time	Throughput per run	Reads per run
High-end instruments					
454 GS FLX + (Roche)	Pyrosequencing	700	23 h	700 Mb	~1,000,000 shotgun, ~700,000 amplicon
HiSeq 2000/2500 (Illumina)	Reversible terminator	2 × 150	High output: ~11 days Rapid run: ~27 h	High output: 600 Gb Rapid run: 120 Gb	High output: 3 billion × 2 Rapid run: 600 million × 2
5500xl W SOLiD (Life Technologies)	Ligation	1 × 75 Frag, 2 × 50 MP	8 days	~320 Gb	1.4 billion × 2
Bench-top devices					
454 GS Junior (Roche)	Pyrosequencing	400	8 h	35 Mb	100,000 shotgun 70,000 amplicon
Ion PGM (Life Technologies)	Proton detection	100 or 200	3 h	100 Mb (314 chip) 1 Gb (316 chip) 2 Gb (318 chip)	400–550 thousand (314 chip) 2–3 million (316 chip) 4–5.5 million (318 chip)
MiSeq (Illumina)	Reversible terminator	2 × 250	27 h	8.35 Gb	6.8 million (LRGC routinely getting >15 M)

2012), oral cavity (Chun et al., 2010), uterine cervix (Smith et al., 2012), male and female urinary tracts (Dong et al., 2011; Siddiqui et al., 2011; Wolfe et al., 2012), asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury (Fouts et al., 2012), and also served as a major sequencing tool in the Human Microbiome Project (Human Microbiome Project Consortium, 2012a,b). However, the popularity of this platform has been impacted in recent years by the emergence of Illumina's next-generation sequencing instruments, as well as Ion Torrent's semiconductor sequencers.

4.3. The Illumina HiSeq2000 platform

In 2006, Solexa released Genome Analyzer (GA). The company was subsequently purchased by Illumina in 2007, which launched the HiSeq 2000 in early 2010, using the same sequencing chemistry as GA. Like Roche 454, HiSeq 2000 also uses sequencing by synthesis following amplification of the input DNA. Briefly, the sample DNA is fragmented and adapters are ligated onto the ends of the DNA fragments. The DNA is selected for size, denatured into single strands and attached to a flowcell, where bridge PCR occurs to form clusters of identical DNA fragments on the cell. The amplified DNA is made single-stranded (Mardis, 2008), and is then sequenced. Nucleotides tagged with fluorescent dyes and a cleavable blocking group are washed over the flowcell; if they are incorporated, the cluster fluoresces with the wavelength associated with that particular nucleotide and the dye and blocking group are cleaved to allow incorporation of the next nucleotide. The Illumina technique has the greatest output and lowest reagent cost: the HiSeq 2000 can reach a throughput of 600 G bases per run and the cost per million bases is only $0.02 (Liu et al., 2012). It is also more accurate than the Roche 454 sequencing platform, since it is not prone to indels at homopolymeric sites, and the read count is steadily improving (Luo et al., 2012). The highly automated library preparation and concentration measurement also greatly reduces hands-on time. Illumina's sequencing technology has been involved in the Human Microbiome Project (Human Microbiome Project Consortium, 2012a), is widely used in taxonomic studies of the microbiome (Lazarevic et al., 2009; Claesson et al., 2010; Gloor et al., 2010; Hummelen et al., 2010), tracking pathogen changes (Harris et al., 2012; Reeves et al., 2011; Grad et al., 2012) and antibiotic resistance, and virulence monitoring (Toprak et al., 2012). The major disadvantages of this method are its relatively short read length and long run time.

4.4. Life Technologies/Applied Biosystems' SOLiD system

Sequencing by Oligo Ligation Detection (SOLiD), acquired by Applied Biosystems in 2006, utilizes a system whereby DNA libraries are prepared, fragmented and ligated to a P1 adaptor with known starting sequences. The fragments are then attached to the magnetic beads and go through emulsion PCR. The resulting PCR product-containing beads are then covalently bound to a glass slide, whereby sequencing is performed by ligating di-base probes, which are fluorescently labeled. SOLiD produces short reads, but due to the di-base ligation method it has a very high accuracy after filtering. The SOLiD 5500xl system, released in 2010, produces 85 bp reads with 99.99% accuracy (Liu et al., 2012). The newest double-flowchip 5500xl W Genetic Analyzer could reach a throughput as high as 320 Gb per run, with the Wildfire upgrade in 2012 further improving its workflow and throughput. The major disadvantages of SOLiD are its short read length, its long run time, and the colorspace mapping of the resulting DNA sequence. The system has been used for whole genome re-sequencing, targeted re-sequencing, transcriptome research including gene expression profiling, small RNA analysis, whole transcriptome analysis and epigenomic studies such as ChIPSeq and methylation (den Bakker et al., 2010; Spanu et al., 2010; Silva et al., 2011; Quast et al., 2012).

4.5. Bench-top sequencing devices

Although high-end sequencing machines can deliver a high throughput and long read lengths, they are also bulky, expensive and usually can only be afforded by large centers. On the other hand, there are also modestly priced, bench-top instruments with throughputs and read lengths decent enough for microbial applications, appearing on the market, and these could be useful for some basic research and clinical applications.

4.5.1. The 454 GS Junior system

The 454 GS Junior was released in early 2010, and has been described as a smaller 454 GS FLX machine, using a similar emulsion PCR and pyrosequencing approach. It has lower throughput than the 454 GS FLX, but also has lower costs for set-up and reactions. Out of all the bench-top sequencing devices, this system produces the longest reads, but has the lowest throughput of the three instruments (35–70 Mb) (Loman et al., 2012b). The cost per base is also at least one order of magnitude higher than the cost for the other two platforms (Loman et al., 2012a). Other disadvantages of 454 GS Junior include its relatively high hands-on time and high error rate in homopolymeric tracts.

4.5.2. Ion PGM from Ion Torrent

The Ion Personal Genome Machine (PGM) was released by Ion Torrent at the end of 2010. Like the MiSeq from Illumina and the 454 GS Junior from Roche, this machine is designed specifically for small labs and clinics. Ion PGM also uses a sequencing-by-synthesis strategy, but unlike other technologies, it directly detects hydrogen ions when bases are incorporated into the growing strand. Each fragment is put in a microwell and dNTPs are washed over the wells. When a dNTP is incorporated by DNA polymerase into the growing strand, a hydrogen ion is released which in turn is detected as a pH change of the solution by an ion sensor, thus identifying the base sequence (Rusk, 2011). This mechanism permits real time sequencing, with a high throughput and relatively short read length of around 200 base pairs. A recent update in May 2013 made the Ion PGM™ Sequencer priced 30% less than the original, and boosted the throughput from 10 Mb to 100 Mb for the Ion 314 chip, 100 Mb to 1Gb for the Ion 316 chip, and up to 2Gb for the Ion 318 chip, giving it the highest throughput per hour with the shortest run time. Its highly automated characteristics also greatly reduce its human labor requirements. The major drawback of Ion PGM is its short read length (100 to 200 bases). Also, like the Roche 454 systems, Ion PGM systems are also prone to error in homopolymeric tracts (Loman et al., 2012a).

4.5.3. Illumina's MiSeq

MiSeq is a bench-top sequencer launched in 2011 that shares many technologies with HiSeq. Due to its smaller single flow cell, it has a shorter run time than that of the HiSeq system. Like HiSeq, it is able to sequence fragments from either end, allowing longer reads, without extra handling during the sequencing step (Loman et al., 2012a). MiSeq has a very low error rate, with the lowest rate of indel mutations of any bench-top sequencer. In addition, many steps can be automated, including library preparation, concentration measurement and template amplification.

4.6. Third generation sequencing techniques

While next generation sequencing techniques are still rapidly evolving, third generation sequencing techniques have brought interesting innovation to the field. In contrast to most second-generation sequencing techniques, which depend on the production of libraries of clonally amplified templates, third-generation sequencing does not amplify the DNA. These techniques have the potential to be less costly, to be less time-consuming, and to have fewer biases from the amplification step, while also capturing their data in real time (Liu et al., 2012).

4.6.1. The Helicos Heliscope Sequencer System

The first single-molecule DNA sequencer, the Heliscope Sequencer System, was launched by Helicos Biosciences in 2008 (Efcavitch and Thompson, 2010). As with Illumina sequencing, the template DNA is immobilized, and synthesis occurs by reversible chain-terminating nucleotides, which are labeled (Bowers et al., 2009; Harris et al., 2008). This system has previously been used to sequence DNA (Pushkarev et al., 2009) and RNA (Ozsolak et al., 2009). Disadvantages of this platform are its high error rates (>5%) and short read length (about 32 bases), making its sequences difficult to analyze (Schadt et al., 2010).

4.6.2. Single-molecule real-time sequencing

Single-molecule real-time (SMRT) sequencing (Eid et al., 2009) developed by Pacific Bioscience (Menlo Park, CA, USA) uses fluorescent dye modified nucleotides, and a zero-mode waveguide (ZMW). The SMRT cell consists of millions of ZMWs, each containing one DNA template. DNA polymerase is fixed at the bottom of ZMWs with a biotin–streptavidin linkage (Korlach et al., 2010), and forms a complementary strand as in normal DNA replication during the reaction, cleaving off the fluorescent dye previously linked to the terminal phosphate of the nucleotide (Liu et al., 2012). This process emits light signals, which is captured by a built-in camera as videos on a real time basis (Timp et al., 2010). This is useful since both the color and intensity are measured, which can give information not only on the sequence but the structure; this technique has thus been used for studying epigenetic base modifications (Branton et al., 2008; Clark et al., 2013; Schadt et al., 2013). The utilization of DNA polymerase's natural abilities enables fast cycle time and very long reads. It also has simple sample preparation, and low reagent costs. SMRT has had other applications to examine biochemical properties of transcription and translation (Travers et al., 2010; Uemura et al., 2010; Lo et al., 2011). The major drawbacks of SMRT include high raw error rates (>10%), low throughput, and expensive and difficult set-up (Stranneheim and Lundeberg, 2012). Recently, Koren et al. reported a combined Illumina/SMRT algorithm to address these limitations, with a read accuracy over 99.9% (Koren et al., 2012).

4.6.3. Oxford nanopore sequencing

Nanopore sequencing enables direct reading of unlabeled DNA by threading it through a nanoscale-sized pore, that is, a nanopore (Song et al., 1996) Biologically, nanopores are usually present as protein channels for ion exchange. Any substances that move through the channel can cause changes of the current across the channel due to their different conductivity, which are monitored and recorded. The identity of each deoxyribonucleoside monophosphate (dNMP) can be determined by their size difference. The advantage of nanopore sequencing is that it can reach a read length more than 5 kbp and a speed of 1 bp/ns (Branton et al., 2008). Since there is no need for fluorescent modification of the bases, it also reduces cost and potential biases. The mechanism of electrophysiological detection also reduces bias related to enzymatic activities (Eisenstein, 2012). Like other third generation techniques, the single molecular sequencing nature of nanopore sequencing greatly decreases the hands-on preparation time including cloning and amplification. Oxford Nanopore Technologies have launched GridION, a commercial sequencing device on an electronic-based platform, and MinION, a disposable portable device for electronic single molecule sensing. The major drawbacks of nanopore technologies include its relatively low throughput, high translocation velocity and the lack of nucleotide specificity (Branton et al., 2008; Venkatesan and Bashir, 2011).

5. Common applications of high-throughput sequencing and bioinformatic tools

High-throughput sequencing techniques produce massive amounts of data and thus to draw any useful conclusions it is necessary to computationally analyze the information. In this section, we discuss the most commonly used methods and tools for targeted amplicon sequencing analysis, shotgun metagenome analysis, and metatranscriptome analysis, and provide some examples of emerging technologies. Fig. 1 summarizes the steps of bioinformatic analyses for each of these experiments.

5.1. Who is here? Targeted amplicon sequencing and analysis of bacterial diversity

Amplifying and sequencing variable regions of highly conserved bacterial genes is a very common way to determine the taxonomic composition of a microbiome. By comparing them to existing databases, one can determine from which organisms the sequences came, and thus determine the bacterial profile and proportions.

5.1.1. Amplicon choice

Typically, the gene encoding 16S rRNA is amplified to analyze prokaryotic taxonomic composition in samples, as it is highly conserved in all prokaryotes. However, since the 16S rRNA gene is about 1550 base pairs long, it is difficult to sequence the entirety of the gene using high-throughput sequencing methods without requiring an assembly step. While techniques such as expectation maximization iterative reconstruction of genes from the environment (EMIRGE) have been developed to assemble the full 16S gene for taxonomic studies (Miller et al., 2011), they are not frequently used, as increasing the length of the sequence studied decreases the depth of coverage for the sequences, making it more difficult to assemble the gene and resolve rare taxa.

Instead, usually one or more of the nine variable (V) regions of the 16S rDNA are amplified and sequenced, using particular sets of primers. However, the variable regions differ between species in different ways—for example, some species can be distinguished in one variable region but not the other. Schloss (2010) reported that the use of different variable regions influences the way OTUs cluster and the reported richness and evenness of communities; he urged caution in analysis of these data, and emphasized that they should not be analyzed in the same way that one would analyze a full-length 16S rRNA gene sequence. Kim et al. (2011) reviewed and analyzed commonly used primers to determine how the use of different variable regions for analysis affects OTU clustering, annotation, and estimates of diversity when applied to sequences available from the Ribosomal Database Project (RDP) database, and concluded that it is easiest to analyze the V1–V3 or V1–V4 regions, as these are more variable and are better represented in the RDP database. Liu et al. (2008) also compared several 16S rRNA amplicon sequencing designs and found that sequencing the V2–V3 regions produced the most accurate results, as determined by modeling short read sequence data based on full 16S rRNA sequences; their paper recommends several primers that can be used for this purpose. Youssef et al. (2009) compared the effects of the choice in variable region on the amount of OTUs detected and found that V4, V5–V6, and V6–V7 sequencing produces the best estimates for species richness in a sample. The Earth Microbiome Project (Gilbert et al., 2010) is standardized to amplify the V4 region, since the primers for that region can detect most *Bacteria* and *Archaea* (www.earthmicrobiome.org).

The choice of which rDNA variable region to amplify depends on a variety of factors specific to the sample and experiment, including the particular bacteria present, whether it is most important to get resolution at a species, genus, or higher taxonomic level and the read length that is afforded by the sequencer. Since all variable regions of the 16S rRNA gene have strengths and weaknesses, some researchers may opt to sequence more than one variable region to get a clearer view of the composition of the microbiome. Further detail into the effects of study design on the results of amplicon sequencing studies can be found in the works by Shah et al. (2010), Soergel et al. (2012), and Mizrahi-Man et al. (2013).

While the 16S rRNA gene is by far the most frequently used gene for studies of community membership and structures, its use has

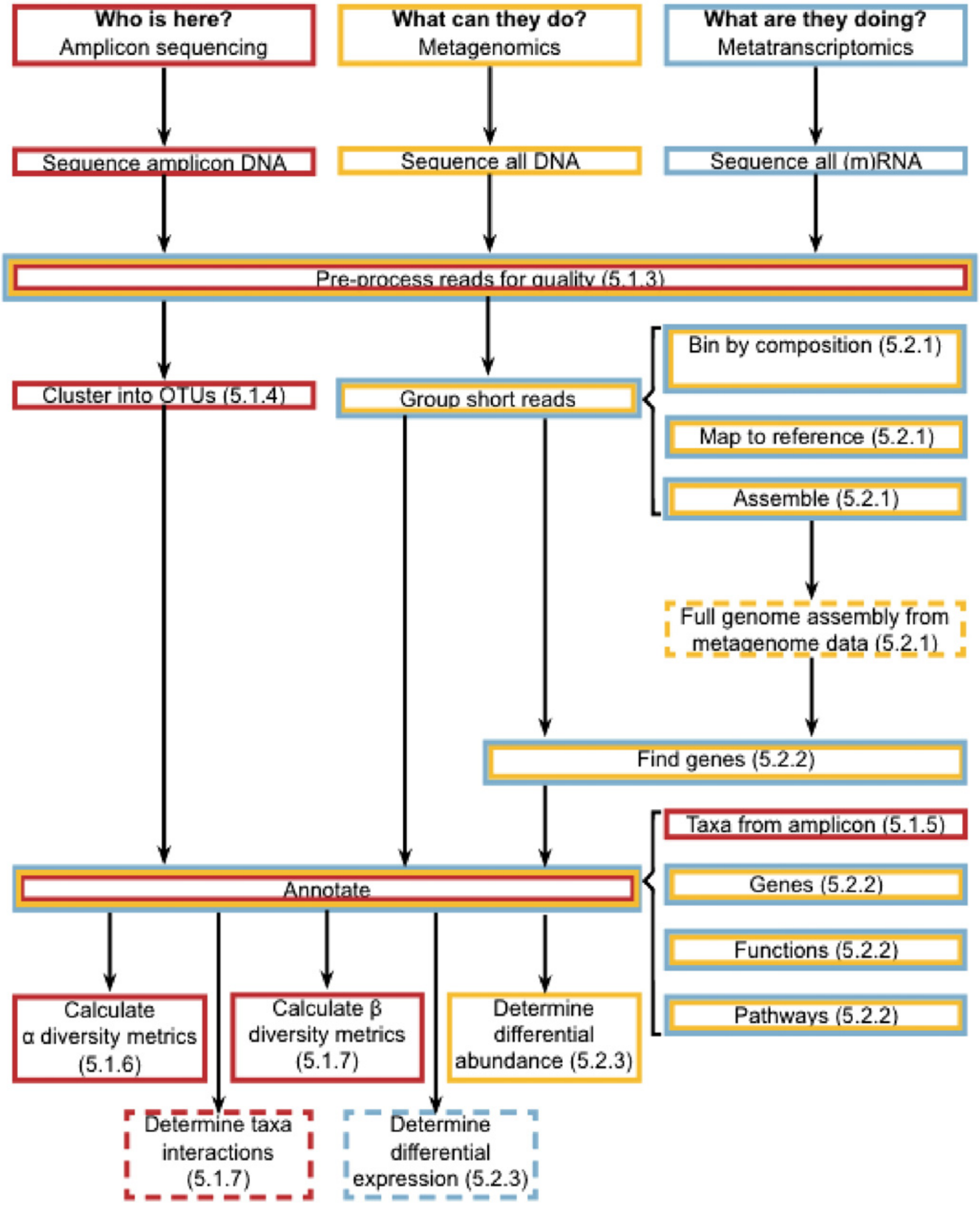
Who is here?
Amplicon sequencing
What can they do?
Metagenomics
What are they doing?
Metatranscriptomics
Sequence amplicon DNA
Sequence all DNA
Sequence all (m)RNA
Pre-process reads for quality (5.1.3)
Cluster into OTUs (5.1.4)
Group short reads
Bin by composition (5.2.1)
Map to reference (5.2.1)
Assemble (5.2.1)
Full genome assembly from metagenome data (5.2.1)
Find genes (5.2.2)
Taxa from amplicon (5.1.5)
Annotate
Genes (5.2.2)
Functions (5.2.2)
Pathways (5.2.2)
Calculate α diversity metrics (5.1.6)
Calculate β diversity metrics (5.1.7)
Determine differential abundance (5.2.3)
Determine taxa interactions (5.1.7)
Determine differential expression (5.2.3)

Fig. 1. Flowchart detailing the steps in bioinformatic analysis of sequence data. Boxes in red represent steps for amplicon sequence analysis, boxes in yellow represent steps for metagenome analysis, and boxes in blue represent steps for metatranscriptome analysis. Boxes with dotted lines represent new analyses for which the computational tools are still under development.

limitations. One main criticism of the technique is that many bacteria have more than one copy of the *rrn* operon, which contains the 16S rRNA gene (Acinas et al., 2004). This is problematic for abundance studies based on 16S sequences, since bacteria with more 16S rRNA genes will be over-represented if this method is used. Furthermore, in some species of bacteria, there are high levels of sequence divergence between the multiple copies of the 16S rRNA gene present in the genome, especially in extremophiles (Acinas et al., 2004). These inflate diversity estimates, as their different rRNA genes make it appear as if they are more than one bacterium. Lastly, some 16S rRNA primers, although considered universal, preferentially bind to some taxa over others, thus over-representing those taxa in experiments.

Other highly conserved genes have been used as an alternative to the 16S rRNA gene. The most commonly used alternative is *cpn*60, a protein-coding gene that encodes type I chaperonins. Previous research has shown that *cpn*60 sequencing has better resolution to distinguish species than 16S rDNA, and it can detect some species that 16S rDNA sequencing cannot (Schellenberg et al., 2009; Hill et al., 2010). However, in species with multiple copies of *cpn*60, the copies tend to be more divergent than 16S rRNA genes are (Links et al., 2012) potentially inflating taxonomic differences similarly to the case for multiple rRNA genes. For other highly conserved genes, such as *rpoB*, no universal primers exist, making these genes difficult to work with for studying the whole microbiome without bias. Another major disadvantage of using alternatives to the 16S rRNA gene is that databases for these genes are less well developed, and thus annotation of the reads is more challenging (see Section 5.1.5 for details on annotation).

5.1.2. Commonly used software packages and pipelines

Amplicon sequencing-based studies are becoming increasingly popular, and so a variety of pipelines and tools have been developed to facilitate analyses of these data. Most of the tools described below can be found in these packages, allowing analysis of sequence data without having to learn many different programs. The most popular pipelines are Quantitative Insights Into Microbial Ecology (QIIME) (Caporaso et al., 2010) and mothur (Schloss et al., 2009), which have all of the most common tools and analyses for amplicon sequencing data.

For less common tools and analyses, the code is often available separately. One useful tool to learn is R, a programming environment that is often used for statistical analyses and visualization of data (www.R-project.org). Many software packages that are useful for analysis of microbiome data have been developed for use in R, such as vegan, which has many tools for assessing diversity (http://CRAN.R-project.org/package=vegan), and phyloseq, which is used to analyze operational taxonomic unit abundance data (see Section 5.1.4) (McMurdie and Holmes, 2013).

5.1.3. Pre-processing of raw amplicon reads

All sequencing methods have some error. Amplicon sequencing analysis is particularly vulnerable to erroneous conclusions caused by incorrectly sequenced reads, as errors in raw sequence reads can lead to erroneously high estimates of bacterial diversity (Kunin et al., 2010). Clustering reads based on sequence similarity counteracts some but not all of these effects (see 5.1.4 for more on read clustering methods) (Kunin et al., 2010). Because of this, pre-processing is often used before further analysis.

Quality-based pre-processing methods are commonly used to filter out low-quality reads, reducing the number of erroneous reads. Previous studies have filtered out reads with ambiguous bases, mismatches to primers, low quality scores, or minimal or no sequence match to the amplified sequence (Huse et al., 2010). Another method that has been used is to use quality scores to trim lengths of low quality sequence from the reads (Kunin et al., 2010), or trimming a fixed number of bases from the ends of reads, where the quality is lowest.

For pyrosequencing methods such as Roche 454, additional tools exist to remove errors due to these techniques' tendency to have insertions and deletions in their sequences at homopolymeric tracts (see Section 4.2). Tools such as AmpliconNoise (Quince et al., 2011) and Denoiser (Reeder and Knight, 2010) analyze the flowgrams from the sequencers to detect potentially erroneous reads.

Another source of sequence artifacts are chimeric sequences created during the PCR amplification step. These form when an aborted PCR product from one sequence binds to and acts as the primer for the extension of another sequence, thus creating a DNA molecule that was made from two templates (Haas et al., 2011). Like errors introduced by sequencing, chimeric sequences can be falsely interpreted as novel amplicons and thus incorrectly increase estimates of bacterial diversity (Huse et al., 2010). Although the risk of chimeras forming is lower when sequencing variable regions, which are not as similar between species, it is still worthwhile to check for them. A number of tools exist to detect and remove chimeric reads, including ChimeraSlayer, which uses a reference set of amplicons (Haas et al., 2011), Perseus, which acts *de novo* (Quince et al., 2011), and UCHIME, which can either use a reference database or act *de novo* to detect chimeras (Edgar et al., 2011).

5.1.4. Operational taxonomic unit (OTU) clustering

It is incorrect to assume that each unique read represents a different species: mutations and sequencing errors can lead to slightly different amplicon sequences within a species. To counteract this, before further analysis, reads are aligned and those that are similar are clustered, forming operational taxonomic units (OTUs). The OTUs are defined by clustering together sequences with under a particular percentage of divergence from each other; as this percentage increases, the number of reads clustered together increases as well and the number of OTUs thus decreases. Clustering of sequences in an OTU is a tradeoff between detecting more species than are actually present by interpreting mutations and sequence errors as separate species, and detecting fewer species than are actually present by clustering similar reads from different species into separate OTUs. Generally, a 97% identity cutoff is typically used as an approximation of species-level resolution; that is, all reads in an OTU made this way have no more than 3% sequence difference from each other (Stackebrandt and Goebel, 1994).

Since OTU picking generally requires comparison of all input sequences with each other, this procedure can take a large amount of computational time to complete. To make this task easier, pre-clustering is usually done. Generally, this is achieved by collapsing all identical reads into one category; however, it is important that the abundance information for these reads is stored for use in subsequent quantitative analyses. Once the reads are pre-clustered, existing tools can be used to pick OTUs.

There are two main approaches to OTU picking. The most commonly used is *de novo* OTU picking, which compares all sequences to each other and clusters them by similarity with no reference to outside sources of data. Cd-hit (Li and Godzik, 2006), UCLUST (Edgar, 2010), ESPRIT-Tree (Cai and Sun, 2011), CROP (Hao et al., 2011), BeBAC (Cheng et al., 2012), and M-pick (Wang et al., 2013) are all tools that can pick OTUs independent of taxonomy. The other approach to OTU picking is taxonomy-based, where tools such as BLAST (Altschul et al., 1997) compare sequences to a database and cluster OTUs based on their sequence similarity to known sequences in the database (see Section 5.1.5 for details on databases and annotation). This approach is less likely to include erroneous sequences in the analysis and immediately annotates the clustered reads; however, it discards reads with sequences that do not match those in the database, which may be novel species. A third option is to combine the two approaches by using taxonomy-based OTU picking first, and then using *de novo* clustering on the remaining reads (Bik et al., 2012).

5.1.5. Annotation

Once OTUs are determined, it is necessary to find out from which species they came. The database used is dependent, of course, on the sequence amplified. For 16S rRNA gene sequences several databases exist. SILVA has sequences for both the small and large ribosomal subunits in prokaryotes as well as eukaryotes; taxonomic classification of rRNA sequences can be done using its SINA tool, available on its website (Pruesse et al., 2007). The Ribosomal Database Project (RDP) contains prokaryotic 16S rRNA sequences, and has a RDP Classifier tool for taxonomic classification of rRNA sequences, as well as several other tools for 16S rRNA amplicon processing and analysis (Cole et al., 2009). GreenGenes also contains prokaryotic 16S rRNA sequences along with a tool to compare sequences to the database (DeSantis et al., 2006). There are fewer databases dedicated to other prokaryotic marker genes. For *cpn*60 sequences, the cpnDB exists to aid annotation (Hill et al., 2004). For other markers, there are few or no dedicated databases; in these cases, it is best to search against a general purpose protein database (see Section 5.2.2).

5.1.6. Describing richness and evenness within samples

Diversity within a sample is often referred to as alpha-diversity. Richness is a description of the number of species present in a sample, while diversity is a measure that combines richness with how evenly different species make up the sample's microbiome. Richness may be determined by the number of OTUs present, but this can also be influenced by the sequencing depth, so metrics such as Chao1 (Chao, 1984) and ACE (Chao and Lee, 1992) can also be used to estimate the minimum number of species present in the sample. Diversity is commonly measured using the Simpson diversity index and the Shannon diversity index; a detailed description of alpha-diversity indices and application in microbiome research can be found in the work by Li et al. (2012).

5.1.7. Comparing samples

Diversity between samples is often referred to as beta-diversity. At its simplest, samples can be compared by which OTUs they have in common, or by visualizing the relative abundance of different OTUs in a heatmap. However, often a more detailed approach is used that may take into account differences in species presence or absence, abundance, or phylogeny. This requires the use of multivariate analyses. Commonly used tools as well as selected emerging techniques are described below; for a more detailed description of the statistical background of many commonly used tools, we direct the reader to the review by Ramette (2007).

To compare samples, first a distance matrix is made that quantifies the differences between samples. This compares which taxa are present, and may or may not consider phylogeny or abundance of different species. Phylogeny-independent metrics treat all sequences equally regardless of their similarity. Sørensen's similarity matrices and the Jaccard index do this qualitatively by determining the distance between samples based on the presence or absence of OTUs, while the Bray–Curtis and Morisita similarity indices do this quantitatively based on the abundance of OTUs in each sample (Chao et al., 2006). These metrics do not take into account differences in sequencing depth and so sequences must be normalized, or alternative methods that consider unobserved species must be used (Chao et al., 2006; Engen et al., 2011). Conversely, phylogeny-dependent metrics consider samples with more similar sequences to be more similar. This is done by first putting the reads on a phylogenetic tree; this can be constructed *de novo* by aligning and comparing samples, or by using a previously constructed reference tree. A good reference tree may more accurately represent the phylogenetic difference between species; however, a *de novo* tree will allow analysis of sequences that do not match anything on the reference tree. The most commonly used phylogeny-dependent similarity metric is the UniFrac distance. The unweighted UniFrac distance considers similarity between sequences in the samples, but does not compare the sequences' abundances between samples, while the weighted UniFrac distance considers both the similarity between sequences and the sequences' abundances (Lozupone et al., 2010).

Once the distances between the samples can be determined, they can be compared by their similarities. One approach is clustering the data by similarity. Methods such as the neighbor-joining method or the unweighted pair-group method with arithmetic mean (UPGMA) can be used to arrange the distance data into clusters of samples, which can then be visualized as a dendrogram to show which samples are most related. Squash clustering is a novel type of clustering that uses the Kantorovich–Rubenstein distance (in microbiome analyses this functions in a similar fashion as a weighted UniFrac distance) to cluster samples by their similarity, taking into account similarity and phylogeny; this method is available in the software package "guppy" (Matsen and Evans, 2013). Unlike other methods, in a dendrogram constructed by squash clustering the distance drawn between internal nodes (points that connect multiple samples) corresponds to the difference between them.

Another approach to comparing samples is the use of principal coordinates analysis (PCoA). Here, the points (*i.e.* the samples) are plotted in multidimensional space so that the distance between them is as close as possible to that in the distance matrix calculated earlier. The direction in this multidimensional space that separates the points the most is the first principal coordinate, the direction that separates the points second best is the second principal coordinate, and so on. Each principal coordinate has a percentage value that is the percent of the variation explained by that dimension. One can plot out the points (samples) where the axes represent the first and second principal coordinates (and for a three-dimensional plot, the third principal coordinate). This allows visualization of the distance data by reducing its dimensionality—points that cluster, in this case, have more similar microbiomes, as determined by the distance matrix determined earlier. Points can be color-coded by metadata (see Section 2) to qualitatively examine if different types of samples cluster or form gradients along any components and biplots can be used to show the abundance of different taxa and in which samples they are often found.

A variation of this method was published recently, called edge principal components analysis (edge PCA) (Matsen and Evans, 2013). Edge PCA uses a phylogeny-dependent distance metric that takes into account the relatedness of taxa, and the resulting principal components correspond to differences in particular clades of bacteria. This method produces an edge PCA plot, which can be interpreted in a similar way to a PCoA plot. It also produces a phylogenetic tree of the organisms in the sample in which organisms that drive a given principal component have their edges color-coded (by whether they drive the principal component in the positive or negative direction) as well as having different thicknesses of the edges depending on how much they drive that principal component. This method is useful if differences in samples are driven by closely related taxa. Edge PCA can be done using the "guppy" software package.

The above methods are mainly exploratory. For more specific analysis, commonly used nonparametric hypothesis testing tools include permutational multivariate analysis of variance (PerMANOVA, also referred to as nonparametric multivariate analysis of variance, NPMANOVA) (Anderson, 2001), analysis of similarity (ANOSIM) (Clarke, 1993), and the Mantel test (Mantel, 1967). These can be used with any distance metric to determine if the microbiota differs significantly between particular groups of samples. Recently, parametric tools have been developed to test hypotheses related to the microbiome, such as those found in the HMP R package. These tools are more powerful than the nonparametric tests but generally require pooling rare taxa (La Rosa et al., 2012).

5.1.8. Describing interactions between bacteria

Since amplicon sequencing and analysis can estimate the composition of different samples, one can theoretically determine interactions between bacteria, for example, which bacteria co-occur and which co-exclude. However, this analysis is deceptively challenging. Since abundances of bacteria determined by these methods are proportional data, they are not independent; if one goes up, one or more others must go down. Standard measures of correlation assume that the data are independent, thus it is not valid to apply these to the proportional data, and doing so may result in spurious correlations (Aitchison, 1982). Because of this, specialized tools must be made for these analyses.

There is currently no tool to analyze co-occurrence or co-exclusion that is frequently used in microbiome analysis, although some tools are under development. Faust et al. (2012) used a novel analytic technique to determine co-occurrence and co-exclusion and analyze the relationships' significance for human microbiome data; however, they did not develop a computational tool to analyze data in the same way. Friedman and Alm (2012) produced a tool, SparCC, that applies the log-ratio transformation (Egozcue et al., 2003) to transform proportional data into independent data and then approximates correlations to find co-occurring and co-excluding bacterial pairs. The problem of determining co-occurrence and co-exclusion based on abundance data is by no means resolved, and developing statistically sound methods that accurately represent and interpret these data is an active area of research.

5.2. What are these bacteria capable of and what are they doing? Shotgun metagenomic and metatranscriptomic analyses

While in amplicon sequencing only one particular gene or region is amplified, sequenced, and analyzed, in shotgun metagenomic analyses, all DNA in a sample is sequenced and analyzed. Not only does this allow determination of which bacteria are present in a sample, but also which genes and gene functions are present. A variation of the method is metatranscriptomic analysis using deep RNA sequencing (RNA-seq), in which all RNA (or, after a purification step, all mRNA) in a sample is sequenced and analyzed. This allows the researcher to determine which genes are being transcribed and at what level.

Metagenomic and metatranscriptomic analyses are much more challenging than amplicon sequence studies for a number of reasons.

One major consideration for these studies is coverage—since these studies sequence whole genomes instead of particular genes, the same sequencing depth yields far lower coverage in metagenomic and metatranscriptomic studies. High coverage is necessary to resolve differences between samples, and so the throughput of the sequencer must be carefully considered when designing these studies. Furthermore, metagenomic and metatranscriptomic analyses have different abundances of bacterial taxa across samples, making assembly of sequences difficult, since assemblers that are not designed for metagenomic data assume a good assembly has even coverage (see Section 5.2.1). This is even more challenging in metatranscriptome experiments, which don't only have different abundances of bacterial taxa, but also have different transcription levels of every gene across samples. In these studies, it is very difficult to determine if differences truly exist, and if they are due to differential abundance of bacteria or differential transcription. To acquire sound results, it is necessary to have a relatively high number of samples to compare, and sufficient coverage in all of them.

The proper method for analyzing metagenomic and metatranscriptomic data is very much an unresolved challenge; however, some tools have been developed to analyze these data and learn how the microbiome acts in response to its environment. Important steps in these analyses and tools that have been developed are described below.

5.2.1. Binning, assembly, and mapping

Once the DNA or RNA is sequenced, it may be useful to put the reads into "bins" with common characteristics that are likely to be from the same or similar organisms. Binning is usually done by analysis of the sequences' composition, since factors such as G + C ratio, di-, tri-, or tetranucleotide frequency, or codon usage tend to differ by bacteria. Binning tools for multispecies samples include TETRA (Teeling et al., 2004), PhyloPythia (McHardy et al., 2006), TACOA (Diaz et al., 2009), PCAHIER (Zheng and Wu, 2010), and AbundanceBin (Wu and Ye, 2011). Some tools consider both sequence composition and similarity to known sequences in databases; these include PhymmBL (Brady and Salzberg, 2009), SPHINX (Mohammed et al., 2011) and MetaCluster (Wang et al., 2012). Binning may be performed before or after assembly, and each binning tool has different requirements—for example, only specific binning tools (*e.g.* AbundanceBin) can use short reads due to limited compositional data within them. Whether or not binning should be done depends on the design of the study—it can be useful to organize and help assemble and/or annotate reads, but it can also be computationally expensive, and there is a risk of mis-binning some reads.

Often, it is useful to combine overlapping reads to form contiguous regions (contigs) before further analysis. This allows detection of longer genes, and can also give information about which genes are adjacent in genomes. The challenge in shotgun metagenomics and metatranscriptomics is that the reads may have come from any region of a genome from any of the organisms in the sample, and assembly runs the risk of assembling chimeras, that is, contigs made of sequences that are not actually contiguous in any genome. This is especially problematic in contigs containing highly conserved genes or mobile elements. Assembly also consumes more computational time as the number of reads increase; samples with a high amount of reads can be quite computationally expensive to assemble. In addition, assembly should not be used if comparing gene abundance between samples, as the process of assembly requires losing the abundance information.

Metagenome assemblers differ from conventional genome assemblers in that they are designed for data containing more than one species, and so they generally have algorithms in place to separate species where possible, decreasing the amount of chimeric contigs constructed. They also tend not to rely on even coverage as a means of verifying assemblies unlike conventional genome assemblers, since coverage is not even in metagenomes because species have different abundances. A variety of assemblers specific to metagenomic data have recently been developed to assemble the best and longest possible contigs taking into account when possible limitations of metagenome assembly. These include Bambus 2 (Koren et al., 2011), Genovo (Laserson et al., 2011), Meta-IDBA (Peng et al., 2011), metagenomic assembly program (MAP) (Lai et al., 2012), and MetaVelvet (Namiki et al., 2012).

Metatranscriptome assembly is a newer field, and can be useful when reference genomes are unreliable or unavailable, and it can aid mapping by making larger contigs to map to a reference. Assembly of metatranscriptome data is very difficult due to high ranges of coverage and very high numbers of short reads. Some assemblers have been developed to facilitate *de novo* metatranscriptome assembly, including Rnnotator (Martin et al., 2010), Trans-ABySS (Robertson et al., 2010), Trinity (Grabherr et al., 2011), and Oases (Schulz et al., 2012).

An alternative to *de novo* assembly is to assemble reads by comparing them to a reference database. However, if the genes in the metagenomic sample are very different from those in the reference database, or if there are genes in the sample that do not exist in the database, this approach will not work well. For further information about comparing read mapping to metatranscriptome assembly, and experimental approaches that use a combination of both, we refer the reader to the review by Martin and Wang (2011). Read mapping tools include BWA (Li and Durbin, 2009), SOAP2 (Li et al., 2009), the Short Read Mapping Package (SHRiMP) (Rumble et al., 2009), Stampy (Lunter and Goodson, 2011), and Bowtie 2 (Langmead and Salzberg, 2012).

A challenge of mapping reads is that some may map to more than one location in the reference sequences, especially if some mismatches are permitted in the mapping. There are some solutions to this problem. If the reference sequences are a collection of genes from different organisms, it may be useful to cluster similar or identical ones using a tool such as Cd-hit (Li and Godzik, 2006) or UCLUST (Edgar, 2010) before mapping. While in these cases it is not possible to determine which transcripts originate from which organisms, it is useful to determine which genes are transcribed on a metagenomic scale. Other options are to allow the mapper to randomly assign the read to one of the sequences the read maps to, or to discard non-uniquely mapping reads in the analysis.

Read mapping in transcriptomics is a much simpler problem to address for prokaryotes than for eukaryotes, which have spliced transcripts. Analysis of eukaryotic metatranscriptome data is beyond the scope of this paper; however, for further information, we direct the reader to the review by Chen (2012).

While the aim of metagenome assembly is generally to construct contigs large enough to assist with gene finding and annotation, it is theoretically possible that in some cases an entire genome can be constructed from metagenomic data. Because of the challenges of assembly from metagenomic data, this is difficult to do and generally requires high coverage and a sample with low diversity. Recent examples of genome construction from metagenomes include the assembly of a *Euryarchaeota* genome (Iverson et al., 2012), the assembly of a *Spartobacterium* genome (Herlemann et al., 2013), and the assembly of the TM6 genome (McLean et al., 2013). Specific tools are beginning to be made for this purpose: recently, the Paired-Read Iterative Contig Extension (PRICE) assembler was developed for targeted assembly from metagenomic data (Ruby et al., 2013). Further methods for facing the challenge of assembling genomes from metagenomic data are under investigation, and better algorithms as well as sequencing technologies are necessary before this method is more widely used.

5.2.2. Gene finding and functional annotation

Whether or not the reads are assembled, before further analysis can be done it is necessary to determine which genes are present and what they do. There are two steps to this process. The first is finding the genes in the sequence fragments. Some tools exist to predict genes *de novo* from properties of the sequence; some examples of tools for gene finding in multispecies metagenomic samples include MetaGeneAnnotator (Noguchi et al., 2008), Orphelia (Hoff et al., 2009), FragGeneScan (Rho et al., 2010), MetaGeneMark (Zhu et al., 2010), Glimmer-MG (Kelley et al., 2012), and metagenomics gene caller (MGC) (El Allali and Rose, 2013).

Because of the challenges associated with assembling metagenomic reads (as described in Section 5.2.1), assembly is often not done for metagenome data. However, gene finding is difficult in metagenome data if the reads are short. Longer sequence reads improve gene finding accuracy; but, the tools with the longest reads, such as Roche 454 sequencing, have high indel error rates in homopolymeric sites (see Section 4.2), which can affect gene finding by causing frameshift errors. Some of the assemblers above include algorithms for finding and correcting these errors. These include FragGeneScan (Rho et al., 2010) and Glimmer-MG (Kelley et al., 2012).

Another approach to gene finding, especially in reads in which gene finding programs do not predict a gene, is matching the read sequences against a protein sequence database (such as the NCBI nonredundant database) to see if any reads match known coding sequences. The downside is that it can only find genes that are similar to genes in the database, and not novel ones.

Once coding sequences are predicted, it is necessary to determine their function. A variety of databases and systems exist to aid in functional annotation of genes and gene segments, including the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), the Clusters of Orthologous Groups (COG) system (Tatusov et al., 2003), Pfam (Bateman et al., 2004), the Conserved Domains Database (CDD) (Marchler-Bauer et al., 2005), SEED (Overbeek et al., 2005), TIGRFAM (Selengut et al., 2007), and eggNOG (Muller et al., 2010). These databases can find domains and classify proteins by function, allowing determination of which functions and pathways are present in the metagenome or metatranscriptome, and in what abundances they are found.

Several computational pipelines exist to simplify annotation in metagenomic studies by talking assembled or raw reads and performing gene prediction and annotation. These include the Metagenome Analyzer (MEGAN) (Huson et al., 2007), the Integrated Microbial Genomes system for metagenomes (IMG/M) (Markowitz et al., 2008), the Rapid Annotation using Subsystem Technology system for metagenomes (mg-RAST) (Meyer et al., 2008), the Simple Metagenomics Analysis Shell for microbial communities (SmashCommunity) (Arumugam et al., 2010), Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) (Sun et al., 2011) and the Human Microbiome Project Unified Metabolic Analysis Network (HUMAnN) (Abubucker et al., 2012). These tools, which differ by the databases used, and the tools used for sequence filtering, binning, and determining function, are useful for metagenomics projects, as they allow a streamlined and previously verified pipeline for analyzing metagenomics data, and relatively easy analysis and visualization of results.

5.2.3. Comparing samples

Many of the tools that exist to compare metagenomic samples are similar to those that compare the phylogenic composition of samples—in the same way one can compare the richness and evenness of taxa, one can analyze the diversity of genes in a sample and determine which samples are metagenomically closely related. In addition, many of the annotation pipelines described in Section 5.2.2 have additional modules for comparing samples.

Comparison of differential gene abundance and expression for specific genes is an active area of research, and new tools are being developed to better model the distribution of reads and analyze the data. As described in Section 5.1.8, sequence data are compositional and cannot be analyzed with statistics that assume independence of the data without first transforming the data. Furthermore, tools that are designed for determining differential abundance and expression often make statistical assumptions about the distribution of reads that are not valid (Fernandes et al., 2013). There is no standardized, accepted way to determine differential gene abundance or expression in metatranscriptomics; however, some tools exist to help identify differentially abundant or expressed genes.

Determining differential abundance is easier than determining differential expression, as the variation comes only from the abundance of species, not the transcription of genes. MetaPath (Liu and Pop, 2011) uses metagenome data and KEGG pathway information (Kanehisa and Goto, 2000) to determine which metabolic pathways are differentially abundant between conditions. ShotgunFunctionalizeR is an R package that determines differential abundance for genes, functions, or pathways between conditions (Kristiansson et al., 2009). The linear discriminant analysis effect size (LEfSe) method finds biomarkers for certain conditions by comparing gene abundance in multiple samples; it can also be used to analyze amplicon sequences (see Section 5.1) and transcriptome data (Segata et al., 2011). By comparing the metagenomic data of different samples, one can determine differences in the metabolic capacity of microbiomes, and identify interactions with the environment and each other.

Using metatranscriptomic data to determine which genes in particular have *differential expression* between environments is a much more challenging problem. Variation in these data sets comes from two sources, the abundance of different taxa and the transcription of different genes. Because of this, there is a wide range of coverage and very high variance. Current methods for determining differential expression from metatranscriptomic data compare expression in different functional gene groups instead of specific genes in order to decrease the variance; however, the variance is still too high for tools that were not designed for multispecies samples. Methods for analyzing these data are in development, but recently some groups have used this approach to gain insight into the behavior of the microbiome. Recently, Maurice et al. (2013) characterized the effects of xenobiotics on the gut metatranscriptome using HUMAnN, LEfSe, and ShotgunFunctionalizeR to analyze differential pathways and functions, while Macklaim et al. (2013) used a new tool for determining differential expression in metatranscriptomic samples, ALDEx (Fernandes et al., 2013), to compare the vaginal metatranscriptome in health and bacterial vaginosis. As the statistical methodologies for metatranscriptomics improve and more metatranscriptome-specific tools are developed, this technique may provide new insight into how the microbiome responds to its environment.

6. Perspectives and conclusions

High-throughput sequencing technologies have improved in output and quality, and have become an indispensable tool for an increasingly wide variety of experiments, including in phylogenetic, diagnostic, and ecological contexts. Through these tools, we can gain insight into the composition, activities and dynamics of a wide variety of microbiomes, helping to elucidate how bacteria interact with each other and their environment.

Updates to laboratory and computational tools are ongoing. In the near future, this will result in sequencers that require less handling (*e.g.* removal of the amplification step) in order to gain more automation and less error, while still maintaining high speeds and throughputs with low costs and error rates. On the computational side, tools are becoming faster and more specialized for multi-species samples, while also being able to investigate new questions such as differential expression from high-throughput RNA sequence data, genome assembly from metagenomic sequences, and interactions between bacteria from amplicon sequence data. Many of these tools have a steep learning curve for the biologist who is not computationally inclined, although specialized workflows are being developed to allow analysis of these data without extensive computational experience. As tools and statistical methodologies are refined, metagenomics and metatranscriptomics will be able to shift the focus of microbiome research from “Who is here?” to “What are they doing?”, allowing deeper insight into how bacteria interact with their environment and each other. Incorporation of metabolomes, nutrients, host genomes, and other metadata will allow the production of an incredibly detailed picture of interactions between the microbiome and its environment, helping us understand the role of prokaryotes in all environments in which they live. In conclusion, the methods outlined in this review are paving the way for a new paradigm of understanding microbiology and the organisms that make up most of the life on Earth.

Acknowledgement

This work was partly funded through the CIHR Vogue team grant.

References

Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Mathangi, T., Henrissat, B., 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. 8, e1002358.

Abulencia, C.B., Wyborski, D.L., Garcia, J.A., Podar, M., Chen, W., Chang, S.H., Chang, H.W., Watson, D., Brodie, E.L., Hazen, T.C., Keller, M., 2006. Environmental whole-genome amplification to access microbial populations in contaminated sediments. Appl. Environ. Microbiol. 72, 3291–3301.

Acinas, S.G., Marcelino, L.A., Klepac-Ceraj, V., Polz, M.F., 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J. Bacteriol. 186, 2629–2635.

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. B Methodol. 44, 139–177.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Aust. Ecol. 26, 32–46.

Arumugam, M., Harrington, E.D., Foerstner, K.U., Raes, J., Bork, P., 2010. SmashCommunity: a metagenomic annotation and analysis tool. Bioinformatics 26, 2977–2978.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., 2004. The Pfam protein families database. Nucleic Acids Res. 32, D138–D141.

Bennett, S., 2004. Solexa Ltd. Pharmacogenomics 5, 433–438.

Berka, J., Chen, Y.-J., Leamon, J.H., Lefkowitz, S., Lohman, K.L., Makhijani, V.B., Rothberg, J.M., Sarkis, G.J., Srinivasan, M., Weiner, M.P., 2010. Bead emulsion nucleic acid amplification. USA Patent 7842457B2.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R., Thomas, W.K., 2012. Sequencing our way towards understanding global eukaryotic biodiversity. Trends Ecol. Evol. 27, 233.

Bowers, J., Mitchell, J., Beer, E., Buzby, P.R., Causey, M., Efcavitch, J.W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G.M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S.M., Steinmann, K., Thompson, J.F., 2009. Virtual terminator nucleotides for next-generation DNA sequencing. Nat. Methods 6, 593–595.

Brady, A., Salzberg, S.L., 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods 6, 673–676.

Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S.B., Krstic, P.S., Lindsay, S., Ling, X.S., Mastrangelo, C.H., Meller, A., Oliver, J.S., Pershin, Y.V., Ramsey, J.M., Riehn, R., Soni, G.V., Tabard-Cossa, V., Wanunu, M., Wiggin, M., Schloss, J.A., 2008. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 26, 1146–1153.

Cai, Y., Sun, Y., 2011. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Res. 39, e95-e95.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., 2010. QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7, 335–336.

Caruccio, N., 2011. Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by *in vitro* transposition. Methods Mol. Biol. 733, 241–255.

Chao, A., 1984. Nonparametric estimation of the number of classes in a population. Scand. J. Stat. 11, 265–270.

Chao, A., Lee, S., 1992. Estimating the number of classes *via* sample coverage. J. Am. Stat. Assoc. 87, 210–217.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T., 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62, 361–371.

Chen, L., 2012. Statistical and computational methods for high-throughput sequencing data analysis of alternative splicing. Stat. Biosci. 5, 138–155.

Cheng, L., Walker, A.W., Corander, J., 2012. Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Res. 40, 5240–5249.

Chun, J., Kim, K.Y., Lee, J.H., Choi, Y., 2010. The analysis of oral microbial communities of wild-type and toll-like receptor 2-deficient mice using a 454 GS FLX Titanium pyrosequencer. BMC Microbiol. 10, 101.

Cieslinski, H., Bialkowskaa, A., Tkaczuk, K., Dlugolecka, A., Kur, J., Turkiewicz, M., 2009. Identification and molecular modeling of a novel lipase from an Antarctic soil metagenomic library. Pol. J. Microbiol. 58, 199–204.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 38, e200.

Clark, T.A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S.W., He, C., Korlach, J., 2013. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing *via* Tet1 oxidation. BMC Biol. 11, 4.

Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143.

Cole, J.R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R.J., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Marsh, T., Garrity, G.M., 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. 37, D141–D145.

Collins, F.S., Morgan, M., Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. Science 300, 286–290.

den Bakker, H.C., Cummings, C.A., Ferreira, V., Vatta, P., Orsi, R.H., Degoricija, L., Barker, M., Petrauskene, O., Furtado, M.R., Wiedmann, M., 2010. Comparative genomics of the bacterial genus *Listeria*: genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11, 688.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., Nattkemper, T.W., 2009. TACOA—taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinforma. 10, 56.

Dong, Q., Nelson, D.E., Toh, E., Diao, L., Gao, X., Fortenberry, J.D., Van der Pol, B., 2011. The microbial communities in male first catch urine are highly similar to those in paired urethral swab specimens. PLoS One 6, e19709.

Duan, C.J., Xian, L., Zhao, G.C., Feng, Y., Pang, H., Bai, X.L., Tang, J.L., Ma, Q.S., Feng, J.X., 2009. Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. J. Appl. Microbiol. 107, 245–256.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194–2200.

Efcavitch, J.W., Thompson, J.F., 2010. Single-molecule DNA analysis. Annu. Rev. Anal. Chem. (Palo Alto Calif.) 3, 109–128.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., Turner, S., 2009. Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

Eisenstein, M., 2012. Oxford Nanopore announcement sets sequencing sector abuzz. Nat. Biotechnol. 30, 295–296.

El Allali, A., Rose, J.R., 2013. MGC: a metagenomic gene caller. BMC Bioinforma. 14, S6.

Engen, S., Grøtan, V., Sæther, B., 2011. Estimating similarity of communities: a parametric approach to spatio-temporal analysis of species diversity. Ecography 34, 220–231.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C., 2012. Microbial co-occurrence relationships in the human microbiome. PLoS Comput. Biol. 8, e1002606.

Fernandes, A.D., Macklaim, J.M., Linn, T., Reid, G., Gloor, G.B., 2013. ANOVA-like differential expression (ALDEx) analysis for mixed-population RNA-Seq. PLoS One 8, e67019.

Ferrer, M., Beloqui, A., Timmis, K.N., Golyshin, P.N., 2009. Metagenomics for mining new genetic resources of microbial communities. J. Mol. Microbiol. Biotechnol. 16, 109–123.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G.G., FitzHugh, W., Fields, C.A., Gocayne, J.D., Scott, J.D., Shirley, R., Liu, L.I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O., Venter, J.C., 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269, 496–512.

Fouts, D.E., Pieper, R., Szpakowski, S., Pohl, H., Knoblach, S., Suh, M.J., Huang, S.T., Ljungberg, I., Sprague, B.M., Lucas, S.K., Torralba, M., Nelson, K.E., Groah, S.L., 2012. Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. J. Transl. Med. 10, 174.

Friedman, J., Alm, E.J., 2012. Inferring correlation networks from genomic survey data. PLoS Comput. Biol. 8, e1002687.

Froehlich, T., Heindl, D., Roesler, A., 2010. Miniaturized high-throughput nucleic acid analysis. European Patent 2224014A1.

Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Levin, J.Z., Livny, J., Earl, A.M., Gevers, D., Ward, D.V., Nusbaum, C., Birren, B.W., Gnirke, A., 2012. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biol. 13, R23.

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., Feng, W., Huson, D., Jansson, J., Knight, R., Knight, J., Kolker, E., Konstantindis, K., Kostka, J., Kyrpides, N., Mackelprang, R., McHardy, A., Quince, C., Raes, J., Sczyrba, A., Shade, A., Stevens, R., 2010. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. Stand. Genom. Sci. 3, 243–248.

Gilles, A., Meglecz, E., Pech, N., Ferreira, S., Malausa, T., Martin, J.F., 2011. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics 12, 245.

Ginzinger, D.G., Godfrey, T.E., Nigro, J., Moore II, D.H., Suzuki, S., Pallavicini, M.G., Gray, J.W., Jensen, R.H., 2000. Measurement of DNA copy number at microsatellite loci using quantitative PCR analysis. Cancer Res. 60, 5405–5409.

Gloor, G.B., Hummelen, R., Macklaim, J.M., Dickson, R.J., Fernandes, A.D., MacPhee, R., Reid, G., 2010. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. PLoS One 5, e15406.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Grad, Y.H., Lipsitch, M., Feldgarden, M., Arachchi, H.M., Cerqueira, G.C., Fitzgerald, M., Godfrey, P., Haas, B.J., Murphy, C.I., Russ, C., Sykes, S., Walker, B.J., Wortman, J.R., Young, S., Zeng, Q., Abouelleil, A., Bochicchio, J., Chauvin, S., Desmet, T., Gujja, S., McCowan, C., Montmayeur, A., Steelman, S., Frimodt-Moller, J., Petersen, A.M., Struve, C., Krogfelt, K.A., Bingen, E., Weill, F.X., Lander, E.S., Nusbaum, C., Birren, B.W., Hung, D.T., Hanage, W.P., 2012. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. Proc. Natl. Acad. Sci. U. S. A. 109, 3065–3070.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulia, D., Tabbaa, D., Highlander, S.K., Sodergren, E., 2011. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. 21, 494–504.

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

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M., 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5, R245–R249.

Hao, X., Jiang, R., Chen, T., 2011. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27, 611–618.

Hardeman, F., Sjoling, S., 2007. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. FEMS Microbiol. Ecol. 59, 524–534.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S.R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., Xie, Z., 2008. Single-molecule DNA sequencing of a viral genome. Science 320, 106–109.

Harris, S.R., Clarke, I.N., Seth-Smith, H.M., Solomon, A.W., Cutcliffe, L.T., Marsh, P., Skilton, R.J., Holland, M.J., Mabey, D., Peeling, R.W., Lewis, D.A., Spratt, B.G., Unemo, M., Persson, K., Bjartling, C., Brunham, R., de Vries, H.J., Morre, S.A., Speksnijder, A., Bebear, C.M., Clerc, M., de Barbeyrac, B., Parkhill, J., Thomson, N.R., 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. Nat. Genet. 44 (413–419), S411.

Heath, C., Hu, X.P., Cary, S.C., Cowan, D., 2009. Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from Antarctic desert soil. Appl. Environ. Microbiol. 75, 4657–4659.

Herlemann, D.P.R., Lundin, D., Labrenz, M., Jürgens, K., Zheng, Z., Aspeborg, H., Andersson, A.F., 2013. Metagenomic *de novo* assembly of an aquatic representative of the Verrucomicrobial class *Spartobacteria*. ASM mBio. 4, e00569-12.

Hess, M., Sczyrba, A., Egan, R., Kim, T.W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M., 2011. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331, 463–467.

Hill, J.E., Penny, S.L., Crowell, K.G., Goh, S.H., Hemmingsen, S.M., 2004. cpnDB: a chaperonin sequence database. Genome Res. 14, 1669–1675.

Hill, J.E., Fernando, W.U., Zello, G.A., Tyler, R.T., Dahl, W.J., Van Kessel, A.G., 2010. Improvement of the representation of bifidobacteria in fecal microbiota metagenomic libraries by application of the cpn60 universal primer cocktail. Appl. Environ. Microbiol. 76, 4550–4552.

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

Human Microbiome Project Consortium, 2012a. A framework for human microbiome research. Nature 486, 215–221.

Human Microbiome Project Consortium, 2012b. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214.

Hummelen, R., Fernandes, A.D., Macklaim, J.M., Dickson, R.J., Changalucha, J., Gloor, G.B., Reid, G., 2010. Deep sequencing of the vaginal microbiota of women with HIV. PLoS One 5, e12078.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. 8, R143.

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

Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. Genome Res. 17, 377–386.

Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., Armbrust, E.V., 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine *Euryarchaeota*. Science 335, 587–590.

Jeon, J.H., Kim, J.T., Kang, S.G., Lee, J.H., Kim, S.J., 2009. Characterization and its potential application of two esterases derived from the arctic sediment metagenome. Mar. Biotechnol. (NY) 11, 307–316.

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kelley, D.R., Liu, B., Delcher, A.L., Pop, M., Salzberg, S.L., 2012. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 40, e9.

Kim, M., Morrison, M., Yu, Z., 2011. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. J. Microbiol. Methods 84, 81–87.

Knierim, E., Lucke, B., Schwarz, J.M., Schuelke, M., Seelow, D., 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One 6, e28240.

Koren, S., Treangen, T.J., Pop, M., 2011. Bambus 2: scaffolding metagenomes. Bioinformatics 27, 2964–2971.

Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D., Adam, M.P., 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat. Biotechnol. 30, 693–700.

Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J., Turner, S.W., 2010. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol. 472, 431–455.

Kristiansson, E., Hugenholtz, P., Dalevi, D., 2009. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics 25, 2737–2738.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P., 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. 12, 118–123.

La Rosa, P.S., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., Shannon, W.D., 2012. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One 7, e52078.

Lai, B., Ding, R., Li, Y., Duan, L., Zhu, H., 2012. A *de novo* metagenomic assembly program for shotgun DNA reads. Bioinformatics 28, 1455–1462.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., Pace, N.R., 1985. Rapid determination of 18S ribosomal sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. U. S. A. 92, 6955–6959.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Laserson, J., Jojic, V., Koller, D., 2011. Genovo: de novo assembly for metagenomes. J. Comput. Biol. 18, 429–443.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osteras, M., Schrenzel, J., Francois, P., 2009. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J. Microbiol. Methods 79, 266–271.

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

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

Li, R., Yu, C., Li, Y., Lam, T., Yiu, S., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966–1967.

Li, K., Bihan, M., Yooseph, S., Methé, B.A., 2012. Analyses of the microbial diversity across the human microbiome. PLoS One 7, e32118.

Links, M.G., Dumonceaux, T.J., Hemmingsen, S.M., Hill, J.E., 2012. The chaperonin-60 universal target Is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. PLoS One 7, e49755.

Liu, B., Pop, M., 2011. MetaPath: identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc. 5, S9.

Liu, Z., DeSantis, T.Z., Andersen, G.L., Knight, R., 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36, e.120.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., Law, M., 2012. Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. 2012, 251364.

Lo, C., Bashir, A., Bansal, V., Bafna, V., 2011. Strobe sequence design for haplotype assembly. BMC Bioinforma. 12 (Suppl. 1), S24.

Loman, N.J., Constantinidou, C., Chan, J.Z., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R., Pallen, M.J., 2012a. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. 10, 599–606.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J., Pallen, M.J., 2012b. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30, 434–439.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., Knight, R., 2010. UniFrac: an effective distance metric for microbial community comparison. ISME J. 5, 169–172.

Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21, 936–939.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., Konstantinidis, K.T., 2012. Direct comparisons of Illumina *vs.* Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7, e30087.

Macklaim, J.M., Fernandes, A.D., Di Bella, J.M., Hammond, J.-A., Reid, G., Gloor, G.B., 2013. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. Microbiome 1, 1–12.

Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., 2005. CDD: a conserved domain database for protein classification. Nucleic Acids Res. 33, D192–D196.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24, 133–141.

Markowitz, V.M., Ivanova, N.N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M.A., Grechkin, Y., Dubchak, I., Anderson, I., 2008. IMG/M: a data management and analysis system for metagenomes. Nucleic Acids Res. 36, D534–D538.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12, 671–682.

Martin, J., Bruno, V., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., Wang, Z., 2010. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. BMC Genomics 11, 663.

Matsen IV, F.A., Evans, S.N., 2013. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. PLoS One 8, e56859.

Maurice, C.F., Haiser, H.J., Turnbaugh, P.J., 2013. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. Cell 152, 39–50.

McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I., 2006. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods 4, 63–72.

McLean, J.S., Lombardo, M.-J., Badger, J.H., Edlund, A., Novotny, M., Yee-Greenbaum, J., Vyahhi, N., Hall, A.P., Yang, Y., Dupont, C.L., Ziegler, M.G., Chitsaz, H., Allen, A.E., Yooseph, S., Tesler, G., Pevzner, P.A., Friedman, R.M., Nealson, K.H., Venter, J.C., Lasken, R.S., 2013. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc. Natl. Acad. Sci. U. S. A. http://dx.doi.org/10.1073/pnas.1219809110.

McMurdie, P.J., Holmes, S., 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Metzker, M.L., 2005. Emerging technologies in DNA sequencing. Genome Res. 15, 1767–1776.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., 2008. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinforma. 9, 386.

Miller, C.S., Baker, B.J., Thomas, B.C., Singer, S.W., Banfield, J.F., 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 12, R44.

Mizrahi-Man, O., Davenport, E.R., Gilad, Y., 2013. Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. PLoS One 8, e53608.

Mohammed, M.H., Ghosh, T.S., Singh, N.K., Mande, S.S., 2011. SPHINX—an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics 27, 22–30.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., Von Mering, C., Doerks, T., Jensen, L.J., 2010. eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. Nucleic Acids Res. 38, D190–D195.

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, e155-e155.

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, 387–396.

Nowrousian, M., 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot. Cell 9, 1300–1310.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 33, 5691–5702.

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M., Milos, P.M., 2009. Direct RNA sequencing. Nature 461, 814–818.

Pathak, G.P., Ehrenreich, A., Losi, A., Streit, W.R., Gartner, W., 2009. Novel blue light-sensitive proteins from a metagenomic approach. Environ. Microbiol. 11, 2388–2399.

Peirson, S.N., Butler, J.N., 2007. RNA extraction from mammalian tissues. Methods Mol. Biol. 362, 315–327.

Peng, Y., Leung, H.C., Yiu, S., Chin, F.Y., 2011. Meta-IDBA: a *de novo* assembler for metagenomic data. Bioinformatics 27, i94–i101.

Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A., Versalovic, J., 2009. Metagenomic pyrosequencing and microbial identification. Clin. Chem. 55, 856–866.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O., 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35, 7188–7196.

Purohit, M.K., Singh, S.P., 2009. Assessment of various methods for extraction of metagenomic DNA from saline habitats of coastal Gujarat (India) to explore molecular diversity. Lett. Appl. Microbiol. 49, 338–344.

Pushkarev, D., Neff, N.F., Quake, S.R., 2009. Single-molecule sequencing of an individual human genome. Nat. Biotechnol. 27, 847–850.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., 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, J.M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Jian, M., Zhou, Y., Li, Y., Zhang, X., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S.D., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65.

Quast, C., Altmann, A., Weber, P., Arloth, J., Bader, D., Heck, A., Pfister, H., Muller-Myhsok, B., Erhardt, A., Binder, E.B., 2012. Rare variants in TMEM132D in a case–control sample for panic disorder. American journal of medical genetics. Am. J. Med. Genet. B Neuropsychiatr. Genet. 159B, 896–907.

Quince, C., Lanzen, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., Sloan, W.T., 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods 6, 639–641.

Quince, C., Lanzen, A., Davenport, R.J., Turnbaugh, P.J., 2011. Removing noise from pyrosequenced amplicons. BMC Bioinforma. 12, 38.

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

Reeder, J., Knight, R., 2010. Rapid denoising of pyrosequencing amplicon data: exploiting the rank–abundance distribution. Nat. Methods 7, 668.

Reeves, P.R., Liu, B., Zhou, Z., Li, D., Guo, D., Ren, Y., Clabots, C., Lan, R., Johnson, J.R., Wang, L., 2011. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. PLoS One 6, e26907.

Rhee, J.K., Ahn, D.G., Kim, Y.G., Oh, J.W., 2005. New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. Appl. Environ. Microbiol. 71, 817–825.

Rho, M., Tang, H., Ye, Y., 2010. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 38, e191-e191.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., 2010. De novo assembly and analysis of RNA-seq data. Nat. Methods 7, 909–912.

Ronaghi, M., 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11, 3–11.

Ruby, J.G., Bellare, P., DeRisi, J.L., 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data, G3: genes, genomes. Genetics 3, 865–880.

Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A., Brudno, M., 2009. SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. 5, e1000386.

Rusk, N., 2011. Torrents of sequence. Nat. Methods 8, 44.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., Smith, M., 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687–695.

Savage, D.C., 1977. Microbial ecology of the gastrointestinal tract. Annu. Rev. Microbiol. 31, 107–133.

Schadt, E.E., Turner, S., Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Genet. 19, R227–R240.

Schadt, E.E., Banerjee, O., Fang, G., Feng, Z., Wong, W.H., Zhang, X., Kislyuk, A., Clark, T.A., Luong, K., Keren-Paz, A., Chess, A., Kumar, V., Chen-Plotkin, A., Sondheimer, N., Korlach, J., Kasarskis, A., 2013. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res. 23, 129–141.

Schellenberg, J., Links, M.G., Hill, J.E., Dumonceaux, T.J., Peters, G.A., Tyler, S., Ball, T.B., Severini, A., Plummer, F.A., 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Appl. Environ. Microbiol. 75, 2889–2898.

Schloss, P.D., 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. PLoS Comput. Biol. 6, e1000844.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28, 1086–1092.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C., 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12, R60.

Selengut, J.D., Haft, D.H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W.C., Richter, A.R., White, O., 2007. TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 35, D260–D264.

Shah, N., Tang, H., Doak, T.G., Ye, Y., 2010. Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. Pac. Symp. Biocomput. 2011, 165–176.

Siddiqui, H., Nederbragt, A.J., Lagesen, K., Jeansson, S.L., Jakobsen, K.S., 2011. Assessing diversity of the female urine microbiota by high throughput sequencing of 16S rDNA amplicons. BMC Microbiol. 11, 244.

Silva, A., Schneider, M.P., Cerdeira, L., Barbosa, M.S., Ramos, R.T., Carneiro, A.R., Santos, R., Lima, M., D'Afonseca, V., Almeida, S.S., Santos, A.R., Soares, S.C., Pinto, A.C., Ali, A., Dorella, F.A., Rocha, F., de Abreu, V.A., Trost, E., Tauch, A., Shpigel, N., Miyoshi, A., Azevedo, V., 2011. Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. J. Bacteriol. 193, 323–324.

Simon, C., Herath, J., Rockstroh, S., Daniel, R., 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. Appl. Environ. Microbiol. 75, 2964–2968.

Smith, B.C., McAndrew, T., Chen, Z., Harari, A., Barris, D.M., Viswanathan, S., Rodriguez, A.C., Castle, P., Herrero, R., Schiffman, M., Burk, R.D., 2012. The cervical microbiome over 7 years and a comparison of methodologies for its characterization. PLoS One 7, e40425.

Soergel, D.A.W., Dey, N., Knight, R., Brenner, S.E., 2012. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J. 6, 1440–1444.

Song, L., Hobaugh, M.R., Shustak, C., Cheley, S., Bayley, H., Gouaux, J.E., 1996. Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science 274, 1859–1866.

Spanu, P.D., Abbott, J.C., Amselem, J., Burgis, T.A., Soanes, D.M., Stuber, K., Loren, Ver, van Themaat, E., Brown, J.K., Butcher, S.A., Gurr, S.J., Lebrun, M.H., Ridout, C.J., Schulze-Lefert, P., Talbot, N.J., Ahmadinejad, N., Ametz, C., Barton, G.R., Benjdia, M., Bidzinski, P., Bindschedler, L.V., Both, M., Brewer, M.T., Cadle-Davidson, L., Cadle-Davidson, M.M., Collemare, J., Cramer, R., Frenkel, O., Godfrey, D., Harriman, J., Hoede, C., King, B.C., Klages, S., Kleemann, J., Knoll, D., Koti, P.S., Kreplak, J., Lopez-Ruiz, F.J., Lu, X., Maekawa, T., Mahanil, S., Micali, C., Milgroom, M.G., Montana, G., Noir, S., O'Connell, R.J., Oberhaensli, S., Parlange, F., Pedersen, C., Quesneville, H., Reinhardt, R., Rott, M., Sacristan, S., Schmidt, S.M., Schon, M., Skamnioti, P., Sommer, H., Stephens, A., Takahara, H., Thordal-Christensen, H., Vigouroux, M., Wessling, R., Wicker, T., Panstruga, R., 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science 330, 1543–1546.

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic note: a place for DNA–DNA reassociation and rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. 44, 846–849.

Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., DeLong, E.F., 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J. Bacteriol. 178, 591–599.

Stranneheim, H., Lundeberg, J., 2012. Stepping stones in DNA sequencing. Biotechnol. J. 7, 1063–1073.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E.E., Ellisman, M., Grethe, J., 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res. 39, D546–D551.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., 2003. The COG database: an updated version includes eukaryotes. BMC Bioinforma. 4, 41.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., Glöckner, F.O., 2004. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinforma. 5, 163.

Timp, W., Mirsaidov, U.M., Wang, D., Comer, J., Aksimentiev, A., Timp, G., 2010. Nanopore sequencing: electrical measurements of the code of life. IEEE Trans. Nanotechnol. 9, 281–294.

Toprak, E., Veres, A., Michel, J.B., Chait, R., Hartl, D.L., Kishony, R., 2012. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat. Genet. 44, 101–105.

Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S., Turner, S.W., 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res. 38, e159.

Turnbaugh, P.J., Gordon, J.I., 2008. An invitation to the marriage of metagenomics and metabolomics. Cell 134, 708–713.

Uchiyama, T., Abe, T., Ikemura, T., Watanabe, K., 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. Nat. Biotechnol. 23, 88–93.

Uemura, S., Aitken, C.E., Korlach, J., Flusberg, B.A., Turner, S.W., Puglisi, J.D., 2010. Real-time tRNA transit on single translating ribosomes at codon resolution. Nature 464, 1012–1017.

Venkatesan, B.M., Bashir, R., 2011. Nanopore sensors for nucleic acid analysis. Nat. Nanotechnol. 6, 615–624.

Voget, S., Steele, H.L., Streit, W.R., 2006. Characterization of a metagenome-derived halotolerant cellulase. J. Biotechnol. 126, 26–36.

Wang, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics 28, i356–i362.

Wang, X., Yao, J., Sun, Y., Mai, V., 2013. M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. BMC Bioinforma. 14, 43.

Waschkowitz, T., Rockstroh, S., Daniel, R., 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. Appl. Environ. Microbiol. 75, 2506–2516.

Wolfe, A.J., Toh, E., Shibata, N., Rong, R., Kenton, K., Fitzgerald, M., Mueller, E.R., Schreckenberger, P., Dong, Q., Nelson, D.E., Brubaker, L., 2012. Evidence of uncultivated bacteria in the adult female bladder. J. Clin. Microbiol. 50, 1376–1383.

Wu, C., Sun, B., 2009. Identification of novel esterase from metagenomic library of Yangtze river. J. Microbiol. Biotechnol. 19, 187–193.

Wu, Y., Ye, Y., 2011. A novel abundance-based algorithm for binning metagenomic sequences using *l*-tuples. J. Comput. Biol. 18, 523–534.

Wylie, K.M., Truty, R.M., Sharpton, T.J., Mihindukulasuriya, K.A., Zhou, Y., Gao, H., Sodergren, E., Weinstock, G.M., Pollard, K.S., 2012. Novel bacterial taxa in the human microbiome. PLoS One 7, e35294.

Yi, H., Cho, Y.J., Won, S., Lee, J.E., Jin Yu, H., Kim, S., Schroth, G.P., Luo, S., Chun, J., 2011. Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. Nucleic Acids Res. 39, e140.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B.W., Blaser, M.J., Bonazzi, V., Booth, T., Bork, P., Bushman, F.D., Buttigieg, P.L., Chain, P.S.G., Charlson, E., Costello, E.K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J.A., Gallery, R.E., Gevers, D., Gibbs, R.A., Gil, I.S., Gonzalez, A., Gordon, J.I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A.L., Kelley, S.T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C.L., Legg, T., Ley, R.E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methe, B.A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K.E., Nemergut, D., Neufeld, J.D., Newbold, L.K., Oliver, A.E., Pace, N.R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D.A., Assunta-Sansone, S., Schloss, P.D., Schriml, L., Sinha, R., Smith, M.I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J.M., Ward, D.V., Weinstock, G.M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J.R., Yatsunenko, T., Glockner, F.O., 2011. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat. Biotechnol. 29, 415–420.

Youssef, N., Sheik, C.S., Krumholz, L.R., Najar, F.Z., Roe, B.A., Elshahed, M.S., 2009. Comparison of species richness estimates obtained using nearly complete fragments and simulated pyrosequencing-generated fragments in 16S rRNA gene-based environmental surveys. Appl. Environ. Microbiol. 75, 5227–5236.

Yuhong, Z., Shi, P., Liu, W., Meng, K., Bai, Y., Wang, G., Zhan, Z., Yao, B., 2009. Lipase diversity in glacier soil based on analysis of metagenomic DNA fragments and cell culture. J. Microbiol. Biotechnol. 19, 888–897.

Zheng, H., Wu, H., 2010. Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. J. Bioinforma. Comput. Biol. 8, 995–1011.

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