Molecular Microbiology: Diagnostic Principles and Practice, 3rd Edition
Edited by David H. Persing et al.
©2016 ASM Press, Washington, DC
10.1128/9781555819071.ch6

Next-Generation Sequencing

CHARLES CHIU AND STEVE MILLER

6

Next-generation sequencing (NGS), otherwise known as deep or massively parallel sequencing, refers to the technological advances in DNA sequencing instrumentation that enable the generation of hundreds of thousands to millions of sequence reads per run. Sequencing of the human genome, which was once a >10-year endeavor by the NIH at the cost of approximately $3 billion (1), can now be done routinely on a single instrument. Rapid advances in technology led to the first-ever FDA clearance of an NGS instrument, the Illumina MiSeq, in 2014 (2), and the development of rapid, miniaturized sequencing devices such as the Oxford Nanopore are ongoing (3). The applications of NGS are wide-ranging and include (i) whole-genome sequencing, (ii) pathogen discovery, (iii) metagenomic/microbiome analyses, (iv) transcriptome profiling, and (vi) infectious disease diagnosis. Here we will focus on NGS technology and the last three applications, because the first two topics are described in detail elsewhere.

OVERVIEW OF NEXT-GENERATION SEQUENCING METHODS

Prior to the 1980s, Sanger sequencing, based on slab or capillary gel electrophoresis of individual DNA fragments (4), was the only available sequencing technology. The technique was laborious, with a turnaround time of 6 to 24 h, and capacity was limited to the sequencing of fragments in 96 or 384 microtiter wells at a time. The approach taken by NGS technologies, on the other hand, is based on preparation of a "library" of DNA fragments to be sequenced (5). The library is typically produced by the clonal amplification of millions of amplified DNA templates at a time, followed by some method to determine the sequences in a massively parallel fashion. The first available NGS system was the Roche 454 pyrosequencing instrument (6), followed by the emergence of "second-generation" systems (7, 8), including the Illumina (formerly known as Solexa) HiSeq/MiSeq/NextSeq, ABI SOLiD, Life Technologies Ion Torrent, and the PacBio RX system. Currently, the Illumina instruments are used in most published NGS studies, including those in the microbiological field, although new "third-generation" platforms, such as those based on nanopore sequencing (9), are now available and being increasingly used.

Charles Chiu and Steve Miller, Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA 94107.

Roche 454 Pyrosequencing and SOLiD Sequencing

Both the Roche 454 instrument and SOLiD systems isolate and amplify single DNA molecules to construct a library for sequencing by a process known as emulsion PCR (Fig. 1A) (10). Emulsification of an oil-water interface leads to the formation of droplets, with each droplet, referred to as a microreactor, containing a bead that is covalently bound to a single DNA template. PCR amplification is then performed across the surface of the bead to generate clonally amplified fragments. For Roche 454 pyrosequencing, the beads are then deposited into individual wells on picotiter plates, and sequencing reagents containing DNA polymerase are added into the wells (Fig. 1A, left). As the complementary strand is synthesized by nucleotide incorporation, pyrophosphate release produces a fluorescent signal that can be recorded by a CCD (charge coupled device) camera for base calling. For SOLiD, after emulsion PCR, the 3′ ends of the DNA template on the bead are modified to permit chemical linkage to the surface of a glass slide (Fig. 1A, middle). When sequencing reagents containing DNA ligase are flowed over the slide, a fluorescent signal is generated that is captured by a CCD camera for base calling. Roche 454 pyrosequencing is classified as sequencing-by-synthesis, because the sequence is being read concurrently with synthesis of the complementary strand by incorporation of fluorescent-labeled nucleotides (11), whereas SOLiD sequencing is classified as sequencing-by-ligation, because sequencing is determined according to the selective mismatch sensitivity of DNA ligase to fluorescently labeled probes (12).

Ion Torrent Sequencing

For the Ion Torrent, which similar to the Roche 454 uses a sequencing-by-synthesis approach, a semiconductor chip is used to detect hydrogen ions released during DNA polymerization (Fig. 1A, right). A library is prepared by emulsion PCR, and amplified fragments are coupled to beads that are individually deposited in sequencing wells. Nucleotides are then added to the chip, with each of the four bases (A, C, T, and G) being introduced one at a time in a predetermined order. As each nucleotide is incorporated during strand synthesis, a hydrogen ion is released that alters the pH value. Changes in pH are converted and measured in voltage values, which are directly proportional to the number of nucleotides that are incorporated during each cycle.

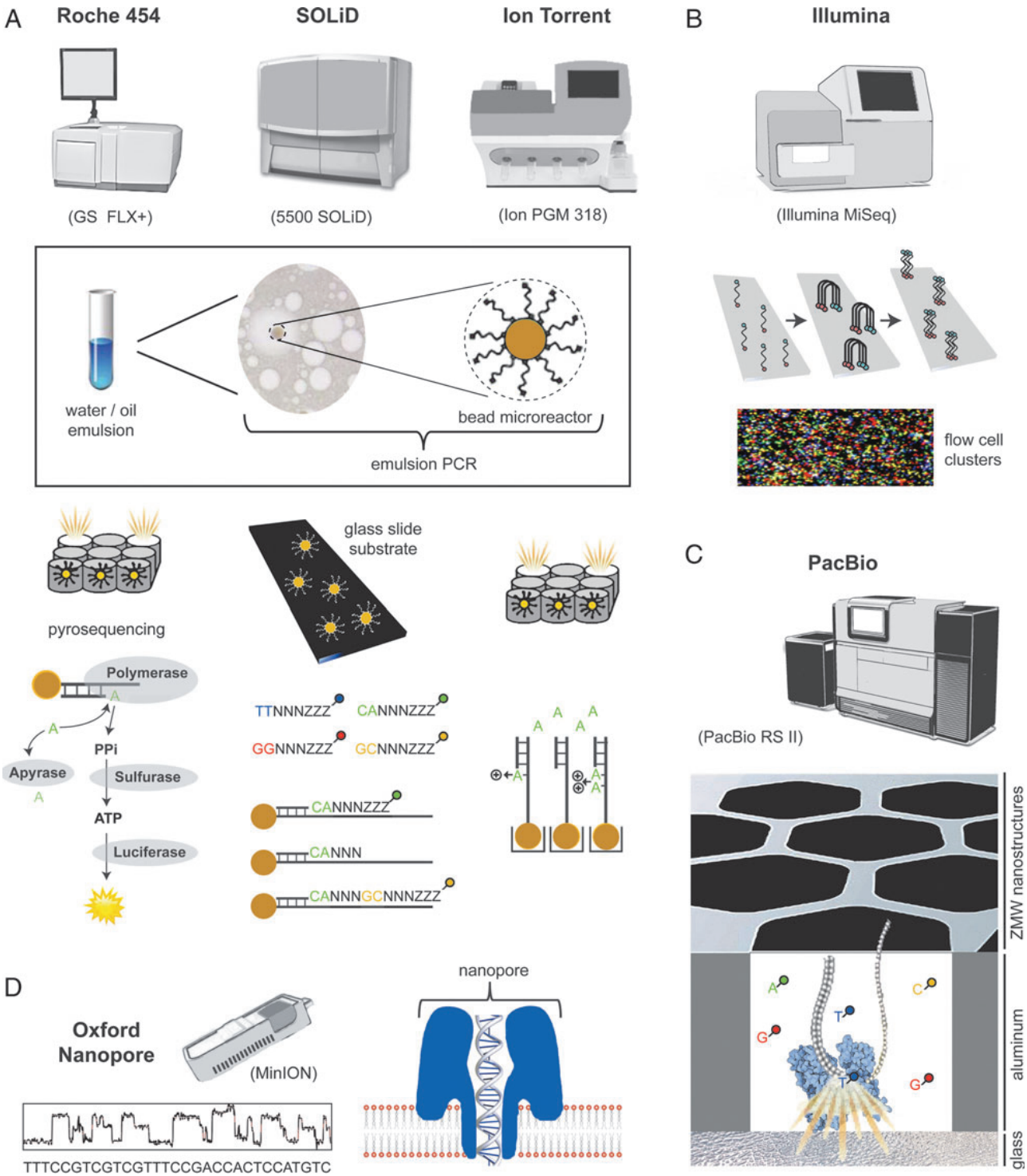

FIGURE 1 Sequencing methods for currently available NGS platforms. (A) The sequencers manufactured by Roche/454 (left), SOLiD (middle), and Ion Torrent (right) all use bead-based emulsion PCR (rectangular inset) in the library generation process, followed by different approaches to fluorescent-based sequencing. (B) Illumina sequencing involves library generation on a flow cell via a sequencing-by-synthesis approach and the imaging of millions of fluorescent flow cell clusters. (C) PacBio sequencing is performed by a DNA polymerase enzyme affixed to a glass substrate in a zero-mode waveguide nanostructure. Each nanostructure generates an individual sequence. (D) Nanopore sequencing, as performed by the Oxford Nanopore MinION instrument, leverages the voltage conductance changes (left) that occur in response to passage of DNA through a nanopore (right), a protein in the lipid-bilayer membrane containing a single hole that allows a single molecule of DNA to pass through.

Illumina Sequencing

The library preparation is simpler than emulsion PCR for Illumina sequencing (Fig. 1B). Two unique primers (adaptors) are attached to the ends of each DNA fragment by ligation, PCR, or transposon switching (Nextera technology) and then affixed to the surface of a flow cell in the form of hairpin loops. "Bridge amplification" using PCR is then performed on the flow cell surface by denaturing the 3′ end of the DNA fragment and replicating the complementary strand. Successive rounds of replication and denaturation by PCR thermocycling result in the generation of thousands of copies of clonally amplified fragments in a tightly circumscribed cluster. Sequencing reagents, including DNA polymerase and a sequencing primer, are then passaged across the flow cell. For each cycle, a single fluorescently labeled nucleotide containing a reversible terminator is added to the complementary strand within each individual cluster in a sequencing-by-synthesis approach. After CCD imaging, cleavage of the fluorescent label permits the next nucleotide to be added. The number of cycles producing the final read length is specified in advance, and sequencing can also be done from both ends (paired-end sequencing) using a second primer to the newly synthesized DNA strand.

PacBio Sequencing

In PacBio single-molecule real-time sequencing technology, individual molecules of DNA template are affixed to the bottom surface of the chip in an optical waveguide called a zero-mode waveguide (Fig. 1C). The zero-mode waveguide creates an illuminated volume within which to observe the incorporation of single nucleotides of DNA. The four nucleotide bases are labeled with different fluorescent dyes and added simultaneously to synthesize the complementary DNA strand. During nucleotide incorporation, the fluorescent tag is cleaved off and a base call is made according to the corresponding fluorescence of the dye (sequencing-by-synthesis). Each single-molecule real-time cell contains approximately 150,000 zero-mode waveguides (13).

Nanopore Sequencing

The sequencers manufactured by Oxford Nanopore use arrays of specialized nanopores that allow a single DNA molecule to pass through at a typical rate of 30 bases per second (range of 0 to 250 bases per second) (14). Current versions of the Oxford Nanopore MinION sequencer, a miniaturized device about the size of a USB stick, contain arrays of 512 nanopores (15), although greater capacity can be achieved with instruments in development including the GridION and PromethION. The current passing through the pores changes in response to the different nucleotide bases as they pass through (Fig. 1D), allowing the sequence to be determined without synthesis, ligation, or other enzymatic steps. Library preparation is simple, but the rate at which a nanopore can capture and sequence a diffusing DNA molecule is limited by concentration, thus requiring relatively high input concentrations of target DNA. At present, error rates in practice are high compared to the more mature sequencing-by-synthesis methods (20 to 40%) (16), but quality consensus sequences can be generated given adequate sequencing coverage.

DIFFERENCES BETWEEN NGS PLATFORMS

Table 1 shows a comparison of the different NGS technologies. The choice of which NGS platform is best suited for any particular application depends on a number of factors including cost, sequencing read lengths, sequencing depth (number of reads per clinical sample) and coverage, and sequencing quality.

Cost

The costs of sequencing have decreased significantly in recent years. Nevertheless, an NGS run is still typically at least an order of magnitude more expensive than that of conventional microbiological assays. Often samples must be individually barcoded and pooled into single runs to decrease costs. As the costs continue to decrease, however, increasing consideration should be given to the cost attractiveness of NGS relative to other recent technologies such as mass spectrometry and microarrays.

Read Length

Longer read lengths are more desirable than shorter read lengths for many applications. For example, in pathogen discovery, longer read lengths facilitate detection of sequences from highly divergent microorganisms such as novel emerging viruses that may be only identifiable on the basis of weak homology in their translated amino acid, rather than nucleotide sequence (17). For metagenomic sequencing, longer reads can also be more accurately classified according to their origin (e.g., human, virus, bacteria, fungus, or parasite), because they are more likely to be uniquely identifying than shorter reads. In addition, longer reads can provide genomic scaffolds that are critical in the *de novo* assembly (joining together of individual reads on the basis of overlapping sequences) of novel microbial genomes for which there is no closely related reference in the database (18). Indeed, many *de novo* assembly approaches combine two technologies: one technology that employs longer but fewer reads (e.g., PacBio, Oxford Nanopore) to enable genomic scaffolding and another technology that employs many more short reads (e.g., Illumina) that can be subsequently mapped onto those scaffolds (3, 19).

Sequencing Depth and Coverage

Sequencing depth and coverage are important parameters for many NGS applications. As a rule of thumb, at least 20× coverage of the genome is generally thought to be needed for accurate *de novo* assembly of a novel organism from short NGS reads (20). In metagenomic "needle-in-a-haystack" applications, a minimum sequencing depth is needed to detect sequences from a target pathogen with high sensitivity amidst a large number of human or animal host background reads (21). The required depth depends on the relative copy number of microbial versus host nucleic acid in the library, with acellular fluids such as serum/plasma, cerebrospinal fluid, and respiratory secretions typically requiring much less sequencing depth at a given level of sensitivity than tissue samples, for which host background sequences are predominant.

Sequencing Quality

Some technologies, such as Roche 454 and Ion Torrent, have difficulty sequencing long homopolymers (22, 23). Other technologies, such as PacBio and Oxford Nanopore, have inherently low individual sequence quality. The low, per-read sequencing quality can be compensated for in PacBio by resequencing the same fragment multiple times

TABLE 1 Comparison of NGS platforms

Platform	Sequencing method	Instrument	Typical read lengths	Accuracy	Throughput (reads per run)	Run time	Instrument cost	Sequencing cost	Key advantages	Key disadvantages
454 Roche	Pyrosequencing	GS FLX+	Up to 700 bp	99.9%	Up to 1 million	20 h	++	+++	Long reads; fast run times	Low throughput; homopolymer errors
SOLiD	Sequencing by ligation	5500 SOLiD	35–50 bp	99.9%	1.0–1.5 billion	1–2 weeks	++++	+	Low cost per base	Very short reads; slow
Ion Torrent	Ion semiconductor	Ion PGM 318	100–200 bp	98.0%	4–5.5 million	2 h	++	++	Fast run times	Homopolymer errors
		Ion Proton I	200–400 bp	98.0%	60–80 million	8 h	+++	++	Fast run times	Homopolymer errors
Illumina	Sequencing by synthesis	HiSeq 2500	50–300 bp	98.0%	0.6–4 billion[b]	6 h to 11 days[b]	++++	+	Highest yield; low cost per base	Instrumentation expensive
		MiSeq	50–300 bp	98.0%	20–30 million	6–40 h	++	+	FDA cleared; low cost per base	Lower throughput
		NextSeq	50–300 bp	98.0%	Up to 800 billion	6–40 h	+++	+	Intermediate yields; low cost per base	Lower throughput
PacBio	Single-molecule real-time (SMRT)	PacBio RSII	10–15 kb	87% or >99.9%[c]	50,000	2 h	++++	+	Long read	Instrumentation expensive
Oxford Nanopore	Nanopore sequencing	MinION	100 bp–10 kb	60–80% or >99%[c]	10,000–50,000	6 h[d]	+	?	Real-time sequencing; portable; long reads	High error rate; low throughput
Sanger[a]	Chain terminator	3730xl	400–900 bp	99.9%	N/A	2 h	++	++++	Long reads; fast run times	Lowest throughput

[a]Not an NGS method; included for purposes of comparison.
[b]Dependent on whether run is rapid-mode or standard-mode.
[c]Individual read or consensus read accuracy.
[d]Can be run until sufficient data are collected; lifetime of flow cell currently 24 to 48 h.

to generate consensus reads (24), while Oxford Nanopore relies on having redundant coverage to compensate for the high error rates (75). The sequencing quality can also vary with length. For example, the quality of Illumina reads deteriorate gradually toward the end of the read (25).

OTHER NGS CONSIDERATIONS

Sample Selection

The NGS approach in microbiology is compatible with a wide range of samples, including clinical human, animal, and even environmental samples, and the choice of sample type is highly dependent on availability and the desired application. When applying unbiased metagenomic techniques that do not rely on specific primers or probes, acellular fluids are preferable to tissues because they have much less host background (21). Metagenomic detection of pathogens is generally less sensitive in whole blood, for example, than in acellular serum or plasma samples. When available, freshly frozen samples are generally superior in quality for NGS applications than formalin-fixed, paraffin-embedded samples or samples allowed to sit at room temperature or 4°C, due to the risk of nucleic acid degradation (26). For applications involving labile RNA such as detection of RNA viruses or transcriptome profiling of mRNA, the use of stabilization reagents at initial sample collection (e.g., PaxGene tubes) should be considered (27). For applications such as infectious disease diagnostics, analysis of more sterile samples such as blood or cerebrospinal fluid is preferred given the increased likelihood of finding a sole causative agent (17), as well as the difficulty in bioinformatics analysis and interpretation of more complex, "environmental" microbial samples such as stool (21). On the other hand, metagenomic and microbiome analyses typically require the presence of a diverse polymicrobial community, such as those found in stool or respiratory secretions. These analyses may not be meaningful for more sterile samples such as blood or cerebrospinal fluid, for which a virome (28), but probably not bacteriome, exists in the healthy state.

For diagnostic NGS, several other considerations need to be taken into account. Collection of noninvasive samples (e.g., sweat, saliva, stool, and urine) is easier than collection of blood or tissue biopsy samples. However, any detected association with NGS is much stronger if made from invasive samples such as tissue biopsy, especially if there is concurrent pathology such as inflammation. Another key consideration is whether to focus on sequencing of library DNA or cDNA generated from RNA. RNA-based NGS is obviously required for RNA virus detection or mRNA transcriptome profiling. For bacterial, fungal, or parasitic identification by 16S/18S rRNA (see below) (29), it may also be preferable to detect transcribed rRNA molecules rather than the rRNA genes to maximize sensitivity, because 10^4 to 10^5 rRNA molecules can be present per microbial cell versus only 1 to 10 copies of the rRNA gene (30). It is also worth noting that RNA-based NGS detection is still capable of detecting DNA-based organisms such as DNA viruses and bacteria by detection of their corresponding host or pathogen mRNA transcripts, respectively. However, RNA is significantly more labile than DNA, and NGS libraries constructed from RNA are also more prone to contamination from exogenous bacterial rRNA from laboratory reagents and the environment (31), which can confound interpretation of the sequencing results.

Disease and Host

Most NGS applications in microbiology are based on direct detection and/or sequencing of microbes. Thus, acute diseases such as febrile illness, which can be associated with high titers of the causative agent (32), are generally more amenable to NGS analysis than chronic diseases. In chronic diseases such as cancer or chronic autoimmune disease, NGS for pathogen detection and discovery relies on the infectious agent still being present at detectable levels in tissue at the time of clinical sample collection. NGS studies in animals can also be more problematic than those in humans, especially if the genome of the animal or a close relative has not yet been sequenced, precluding computational host subtraction approaches to simplify the data analysis (see below).

Sample Preparation Methods

Clinical and environmental samples for NGS are prepared in a series of steps beginning with nucleic acid extraction followed by library preparation (+/– sample barcoding) and loading onto the instrument for sequencing (Fig. 2, left). Initial sample preparation and nucleic acid extraction methods vary depending on the assay type, sample matrix, and pathogen type being detected. Methods to reduce host background or enrich for microbial sequences include ultracentrifugation, nuclease treatment (either pre- or postextraction), and probe-based enrichment. Ultracentrifugation allows for enrichment of virus-like particles, enhancing viral detection (33, 34). Treatment with DNase or RNase will enrich for RNA or DNA targets, respectively, and can substantially reduce host background (33, 34). Probe-based enrichment can be performed using a panel of targets to recover specific organisms from low-titer samples (35).

Sample Barcoding

To multiplex analysis of specimens in a single NGS assay, each sample can be barcoded by adding a short oligonucleotide tag 6 to 12 base pairs (bp) in length to each end of the DNA molecule. Barcoded libraries are then mixed and sequences classified bioinformatically based on the sequenced barcode. To reduce barcode switching, the barcodes are designed to be different by more than one base pair change in a single sequencing run. The use of Hamming code-based designs can preserve minimal distance (in number of base pair changes) between barcodes, and also enable error correction (36). Separate barcodes can be attached to each end of the sequencing target (dual-index barcoding), and barcodes can be rotated over time, reducing the risk of carryover contamination.

Library Preparation Methods

Once the samples have been prepared and nucleic acid extracted, the library is constructed. Each instrument method requires an optimal input amount, which can be generated by preamplification. The final library is generated using emulsion PCR or sequencing adaptor ligation specific to each method (Fig. 1). Library quality control is performed by determining the concentration and size distribution using capillary electrophoresis or real-time PCR. Individually prepared libraries with different barcodes can be pooled for sequencing on a single run, depending on the desired number of sequences per sample.

Contamination

Due to the high sensitivity offered by sequencing large numbers of reads, NGS approaches are extremely vulnerable to

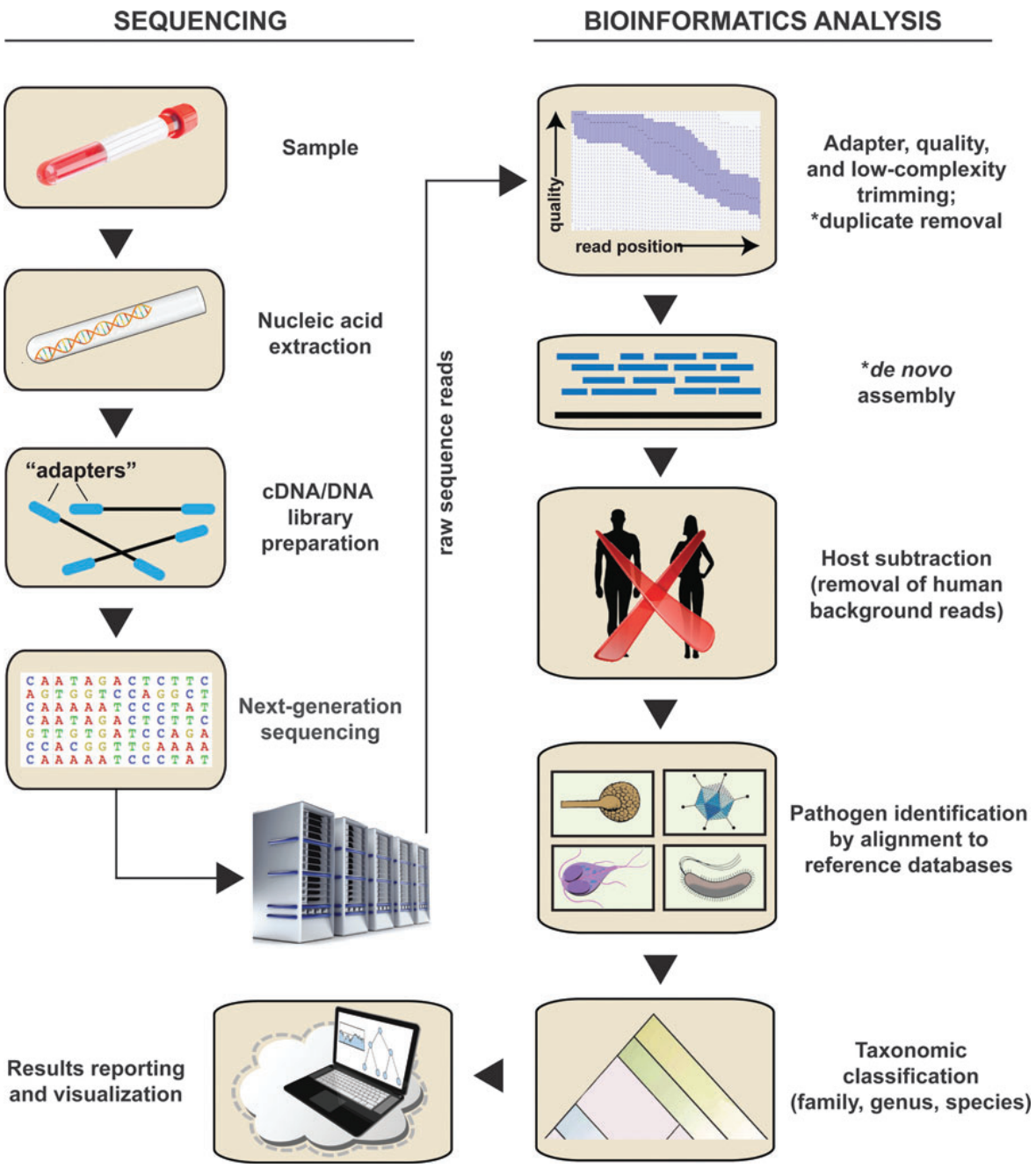

FIGURE 2 Schematic overview of an NGS pipeline. Sample processing for NGS involves a stepwise process of nucleic acid extraction, library preparation, and sequencing on a dedicated instrument (left). Following generation of raw data, bioinformatics analysis of metagenomic or microbial NGS data includes preprocessing, *de novo* assembly, host subtraction, pathogen identification, taxonomic classification, and results reporting/visualization (right). The asterisks denote optional steps in the procedure.

contamination (31). There are multiple points where contaminating organisms or nucleic acid may be introduced into the system, including sample collection, sample processing, and on the sequencing instrument. NGS traditionally requires handling of libraries in an open environment with multiple steps, so amplified material may cross-contaminate samples prepared simultaneously or during subsequent sequencing runs. Reagents used for NGS analysis may be contaminated with microbial nucleic acid, because it is difficult to completely remove DNA from recombinant enzymes. Even commonly used supplies can harbor microbial contamination, such as silica-based DNA purification columns containing what is now thought to be an algal virus (37). Instrument carryover can also occur and is seen both within a run and between runs. Finally, cross-contamination of barcoded samples that are multiplexed in a single run can occur, especially if an individual sample contains a high titer of a specific microbial agent, reads from which can "spill over" into adjacent barcoded samples. This can be mitigated, but perhaps not eliminated entirely, with the use of dual-indexed barcodes at both ends of library amplicons. Thus, careful handling, unidirectional

sample flow, proper quality control, and careful measurements of levels of background contamination are necessary to reduce the risk of false-positive identifications using NGS.

Quality control of NGS reagents is yet another key step to minimize false-positive identifications, particularly with low-input samples having minimal titers of target nucleic acid. Despite efforts to produce ultra-pure reagents (38), there will likely never be assurances that reagents are truly nucleic acid–free. Thus, each new lot of reagents should always be tested with negative controls, and laboratories need to understand the expected frequency and distribution of reagent-derived contaminating sequences and establish appropriate threshold levels of detection to avoid false-positive calls. Common laboratory water supplies often contain bacterial DNA from organisms such as environmental *Burkholderia* and *Ralstonia* species (39, 40), making it difficult to distinguish a true positive identification from background levels of contamination. Also, aerosolized nucleic acid has the potential to contaminate sample hoods and can become a major component of libraries prepared in the hood, requiring extensive cleaning. For certain NGS applications such as pathogen discovery and clinical detection of unusual or unexpected agents, it may be desirable to confirm the results using different extraction methods or reagents or even running the NGS assay in separate laboratories. Confirmation using an orthogonal method such as specific PCR testing from the original sample may also be necessary to exclude the possibility of contamination.

NGS BIOINFORMATICS WORKFLOWS

The sheer number of NGS reads generated by existing instruments and rapid increases in sequencing capacity pose a major computational challenge for analysis of NGS data. A number of bioinformatics software choices are now available, both commercial and open source. For the most part, some degree of computational expertise is needed to take full advantage of these algorithms and workflows, although user-friendly options for NGS analysis exist, such as Geneious (41) and Galaxy (42). Although the details can vary significantly, a computational pipeline for processing and analyzing NGS data follows a general schema (Fig. 2, right). First, sequencing reads are preprocessed by trimming of adapter, low-quality, and low-complexity sequences, with optional removal of duplicate reads. With the exception of host transcriptome profiling using RNA-Seq (43, 44), which deals with alignment and classification of human mRNA genes and isoforms (see below), the next step is to computationally subtract background host sequences (45). For human clinical samples, NGS reads are aligned to the human genome and then removed from the dataset, which decreases the number of remaining reads that need to be analyzed using more computationally intensive downstream algorithms. Next, microbial sequences are identified by alignment to pathogen-specific reference databases such as the National Center for Biotechnology Information (NCBI) bacterial or viral RefSeq databases. Specialized applications such as 16S rRNA sequencing for microbiome analysis classify reads on the basis of alignments to the rRNA gene sequences in the Ribosomal Database Project database (46). Recent advances in the speed and efficiency of alignment algorithms have even made simultaneous alignment to all nucleotide sequences in the NCBI nucleotide (NT) database, including all of GenBank NT (~160 gigabases of sequence as of February 2014; ftp://ftp.ncbi.nlm.nih.gov/genbank/release.notes/), computationally feasible (21).

In addition to sequence alignment, either *de novo*, seed-based, or mapped (using a discrete reference) assembly can be performed to join NGS reads together into contiguous sequences (contigs) and recover partial or even full genomes (47). With metagenomic data, the use of an ensemble method that partitions the data beforehand and combines the use of multiple assembly algorithms may be preferable to maximize contig lengths (48). Translated nucleotide alignment to a protein database or remote homology detection using hidden Markov models (49) can be useful in identifying sequences corresponding to highly divergent pathogens, such as novel viruses. Finally, for NGS applications such as infectious disease diagnosis, precise taxonomic classification of reads to the species level is a necessary step in the analysis (50–52). For example, it is often clinically relevant to be able to distinguish *Staphylococcus* species (e.g., *Staphylococcus aureus* versus coagulase-negative staphylococci) or influenza subtypes (e.g., influenza A [H3N2] versus 2009 pandemic influenza A [H1N1]).

Especially for clinical applications, the development of visualization tools and cloud-computing-compatible platforms will be critical in providing interpretation and context to the NGS data analysis. Software that is user-friendly and produces results that are understandable by microbiologists who lack bioinformatics expertise is greatly needed to enable communication of accurate NGS results to clinicians. A key aspect of NGS for clinical microbiology laboratories will also be not only standardization of the bioinformatics analysis workflows but also standardization of the reference databases. There is currently no consensus as to what standard reference databases will be needed for microbial NGS applications and who would be responsible for developing and maintaining such a database. Nevertheless, working groups consisting of the FDA, NCBI, CDC, and other institutions have been formed to discuss and implement standardized microbial reference databases for NGS (75) as part of a larger effort to ensure the quality of next-generation sequencing in clinical laboratory practice (53).

NGS APPLICATIONS

Amplicon Sequencing

NGS is suitable for sequencing of PCR amplicons in a massively parallel fashion. Applications include determination of minority sequence variants or viral quasispecies and targeted metagenomic analysis. NGS analysis of specific amplicons can deconvolve multiple species in mixed infections, allowing each component to be recognized, whereas Sanger sequencing requires the majority sequence to comprise at least 75% of the total.

Universal Bacterial Identification by 16S PCR

Although the 16S small rRNA gene is found in all bacteria and is highly conserved, the presence of hypervariable regions in the gene sequence allows it to be useful for specific diagnostic identification to the genus and even species level (54). The 16S rRNA gene is 1.5 kB in length and consists of nine hypervariable regions flanked by highly conserved regions. Universal bacterial primers targeting

the conserved regions enable amplification and subsequent sequencing of the hypervariable regions.

A clinically validated assay based on 16S rRNA PCR followed by NGS has been shown to be useful for universal diagnostic identification of bacterial pathogens directly from clinical samples (55). This approach has the advantage of not relying on "gold-standard" culture-based identification, which requires that organisms are capable of growing and replicating *in vitro*. Such an assay based on 16S rRNA PCR would be able to detect fastidious or slow-growing organisms or those rendered nonviable by prior antibiotic treatment or processing (e.g., formalin-fixed paraffin-embedded tissue samples). With the sequencing depth provided by NGS, the presence of even low-titer microorganisms in a highly diverse, polymicrobial sample can potentially be identified. The 16S rRNA gene is also used in most environmental metagenomic studies (56), because it can reveal the phylogenetic relationships among complex bacterial populations at very high resolution. Other targets in bacteria that have been used for these applications include the 23S gene and the intergenic spacer region located between 16S and 23S (57).

UNIVERSAL EUKARYOTIC IDENTIFICATION BY 18S AND/OR ITS PCR

Analogous to the 16S rRNA gene in bacteria, eukaryotic microorganisms that lack a backbone (nonchordate eukaryotes) such as fungi and parasites are identifiable on the basis of 18S or 28S rRNA sequences (58). For fungi, the internal transcribed spacer (ITS) regions can also be used. The hypervariable regions within these sequences can be used to classify fungi and parasites to the species level, and NGS can be readily used for metagenomic analysis as well as provide high sensitivity for detecting low-titer organisms in mixed infections. Because the 18S and 28S rRNA genes are also found in high-order eukaryotes such as animals and humans, inadvertent host background amplification can be significant, generally requiring higher sequencing depths for successful microbial identification.

Pathogen versus Commensal

Many microorganisms are commensals that colonize various body niches of their host and are only associated with disease in the setting of invasion. For instance, fungi such as *Malassezia* spp. and bacteria such as *Staphylococcus* spp. and *Propionibacterium acnes* colonize the skin of healthy adults (59). Therefore, the presence of microbial sequences from nonsterile body sites needs to be interpreted in the context of the infectious disease being studied. A positive detection from a sterile body site is more likely to be associated with true infection but requires differentiation from potential contamination. Also, microbial nucleic acid does not necessarily indicate the presence of live microorganisms but could simply indicate prior colonization. Assessment of the patient's symptoms and clinical presentation, along with the sequencing results, is necessary to determine the pathogenic significance of any microorganisms detected by NGS analysis.

METAGENOMIC AND MICROBIOME ANALYSES

Metagenomic sequencing is targeted (e.g., 16S) or shotgun sequencing of clinical or environmental samples and is now being largely performed by NGS given the depth of coverage that can be achieved. The microbiome, the totality of microorganisms that reside in diverse niches of the human body (60), can be assayed using metagenomic sequencing. The Human Microbiome Project, started in 2008, used 16S sequencing to profile microbial communities at different body sites and thus characterize the baseline microbiome responsible for the maintenance of human health (61). 16S metagenomic or microbiome sequencing can now be routinely performed using customized workflows such as QIIME to classify reads into operational taxonomic units and assess sample diversity (62). Similarly, 18S/ITS or shotgun metagenomic sequencing can be done to analyze fungi for high-resolution species identification and overall profiling of complex microbial communities.

TRANSCRIPTOME PROFILING

Transcriptome profiling by NGS, otherwise known as RNA-Seq, has many applications to microbiology. Transcriptome profiling by NGS is the sequencing of all of the mRNA molecules from either the host or the microorganism to obtain a global view of the gene expression pattern in a clinical sample (43, 44). For full coverage of the human transcriptome, approximately 30 to 50 million short reads are needed. Previously, only microarrays were available to conduct comprehensive gene expression analyses. By transcriptome analyses of the human host response to infection, microarray-based methods have proven effective in the diagnosis of staphylococcal bacteremia (63), active versus latent tuberculosis (64), and acute respiratory infections such as influenza (65). RNA-Seq using NGS has been shown to be more sensitive for detection of low-abundance transcripts, with a broader dynamic range in detecting fold-changes in gene expression at the cost of greater complexity of analysis and current lack of standardization (66, 67).

Transcriptome profiling of the microorganism is also possible, either in pure experimental cultures *in vitro* or directly from clinical samples (68). The data from mRNA gene expression is compared to that from the DNA genome. Microbial transcriptional profiling may yield insights into the overall activity of the organisms (latent versus active metabolism), growth characteristics (aerobic versus anaerobic growth), or expression of resistance and virulence elements.

INFECTIOUS DISEASE DIAGNOSTICS

There is much excitement about the potential of NGS to cause a paradigm shift in microbiology by complementing or even replacing existing diagnostic tests in the clinical laboratory. Metagenomic NGS in particular is promising for diagnosis because this unbiased approach does not target any individual microbial agent but, rather, identifies any and all potential pathogens simultaneously on the basis of sequence homology (17, 21). The capacity of metagenomic NGS to generate clinically actionable data was recently demonstrated in its use to diagnose a case of neuroleptospirosis in a critically ill child that had eluded all conventional diagnostic testing for 4 months (69). Once the diagnosis was made, appropriate targeted therapy resulted in a prompt recovery and cure.

However, translation of NGS assays from research tools for microbial characterization, pathogen discovery, and

epidemiological investigation to actionable clinical diagnostic tests introduces a number of new challenges. Reproducibly generating acceptable libraries from a variety of specimen types that vary by orders of magnitude in human and microbial nucleic acid content is difficult and currently requires multiple parallel strategies. Samples with low organism loads may require pathogen enrichment or amplification, while tissues with high human DNA content may need host subtraction techniques. Each additional step must be controlled for quality and has the potential to introduce contamination, so it is preferable to minimize processing steps where possible. To date, we are unaware of any universal library preparation protocol that can be used to detect all pathogen types in clinical samples with high sensitivity and specificity. One potential workaround is to bias the detection for specific pathogens using a targeted probe enrichment or amplification panel approach followed by NGS instead of relying on shotgun metagenomic NGS for diagnosis (35).

Even with technical hurdles cleared, it remains to be seen whether NGS allows for improved efficiency when compared to conventional clinical diagnostic testing. Certainly the promise of enhanced breadth of detection and genomic characterization is compelling, since it could allow for more personalized medicine and individualized treatment regimens. NGS-based analysis of the host transcriptome response using RNA-Seq may provide complementary information that can be used to guide or modify the approach to patient management and treatment. Furthermore, many studies are now describing how the human microbiome and pathogen genotype influence disease progression, but our knowledge in this area is far from complete. However, we expect that ongoing findings and insights from NGS in microbiology will enable a more comprehensive perspective regarding health and disease states and eventually lead to treatments targeted to specific aberrations in the host and microbial genomic profile.

CLINICAL VALIDATION

Clinical validation of a metagenomic or targeted NGS assay is a substantial undertaking, designed to demonstrate acceptable performance characteristics for an essentially unlimited number of pathogen targets and sequence variants. The assay should be shown to be significantly robust with valid limits of detection, accuracy, specificity, and reproducibility (70). Here, the traditional approach to single-analyte validation fails, because it is impossible to confirm the presence or absence of all possible organisms using standard reference methods. Instead, a validation approach that aims to identify and reduce potential sources of error in the test may be a practical alternative. For infectious disease NGS, this can be done using representative pathogen types in clinical matrices of interest, along with a thorough analytic evaluation to identify error-prone steps and introduce specific quality controls designed to detect errors when they occur. Controlling for sources of contamination is particularly important and should be addressed in the workflow and implementation of routine internal and external controls. Additionally, the NGS data analysis pipeline and reference databases will need to be separately validated. Establishment of curated standard reference databases will likely to be needed, in parallel with the use of additional bioinformatics analysis and review steps to identify misannotated or incomplete database entries. Finally, the reports must be interpretable by clinical microbiologists, be understandable to treating physicians, and provide clinically relevant and actionable results.

REGULATORY AND OTHER CONSIDERATIONS

Currently, no NGS assays for infectious disease diagnosis have been approved by the FDA, though clinical laboratories are starting to offer them as laboratory-developed tests. Proposed regulatory changes initiated by the FDA would likely establish a mechanism for review of newly developed NGS assays, and additional requirements may be instituted in the future to ensure that these clinical tests are safe and efficacious (71). While clinical laboratories are familiar with the requirements under the Clinical Laboratory Improvement Amendments for test validation, quality control, and proficiency testing, they do not typically establish *de novo* clinical utility for these assays. Indeed, the clinical trial design, outcome measures, and statistical confidence needed to demonstrate clinical utility are unknown. It will likely take a coordinated effort between academia and industry as well as stepwise guidance by the FDA to bring NGS for infectious disease to regulatory approval.

The validation of bioinformatics pipelines and databases is another challenge that is beginning to be addressed, but a suitable solution is not yet available. Analysis tools are being continually refined for speed and accuracy, but there is no standardized method to compare them or benchmark their performance. Curated databases typically have a limited number of microorganisms represented, and large public databases such as NCBI NT contain many misannotated sequences that could lead to erroneous results and interpretation (72). Curated 16S ribosomal databases are available for bacterial amplicon sequencing, but databases for other targets and whole-genome sequences are less well characterized. On the other hand, if a standardized reference database is successfully established, it is possible that it can be used as a sole comparator to establish the performance of an NGS assay, forgoing the need for traditional confirmation by orthogonal testing. Given the risk of contamination with metagenomic NGS, multisite evaluation would likely be a requirement for regulatory approval. Development of a panel of representative microorganisms that would function as microbial reference standards, under way at National Institute of Standards and Technology, would also likely be needed for NGS validation (73).

Due to the complexity and data storage requirements for high-throughput NGS analysis, cloud computing and remote storage are attractive options. However, demonstrating HIPAA (Health Insurance Portability and Accountability Act) compliance may be difficult, and there is a risk of data loss during transfer or storage. Most clinical laboratories are unfamiliar with the establishment or maintenance of large computational servers and databases, and the requirements for remote systems are not always clear. The use of bioinformatics tools for NGS analysis and interpretation of results are also not part of the routine skill set of most microbiology laboratories, so simpler graphical visualization interfaces and additional training in bioinformatics may be needed to enable these tests to be more broadly accessible to laboratory personnel. Standards will need to be established for the storage of clinical and technical metadata in addition to the sequence data. Finally, advances in health information technology and electronic medical records software will be required to determine how best to incorporate NGS information into the patient medical record.

CONCLUSIONS AND PERSPECTIVE

NGS assays hold great promise for the broad identification and genomic characterization of infectious disease pathogens. A variety of NGS technologies are now available, each with specific advantages and disadvantages. Sequencing assays incorporating pathogen detection, microbiome analysis, and host transcriptome profiling may lead to more personalized treatment approaches in the future. Several technical hurdles remain to be overcome prior to routine use, including optimal library preparation techniques for different microorganism and sample types, choice of bioinformatics pipelines, and suitable reference databases for comparison. The pathologic significance of microbial detection requires interpretation within the clinical context and may need additional confirmatory testing, particularly for detection of unexpected and/or novel agents. A multifaceted approach involving clinical and research laboratories, bioinformatics scientists, biotechnology companies, and regulatory agencies will likely be needed to take advantage of the large and complex sequence datasets that are currently generated by NGS analysis.

REFERENCES

1. **Hayden EC.** 2014. Technology: The $1,000 genome. *Nature* **507:**294–295.
2. **Collins FS, Hamburg MA.** 2013. First FDA authorization for next-generation sequencer. *N Engl J Med* **369:**2369–2371.
3. **Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, Wain J, O'Grady J.** 2015. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat Biotechnol* **33:**296–300.
4. **Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M.** 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265:**687–695.
5. **Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P.** 2014. Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56:**61–64, 66, 68 passim.
6. **Ronaghi M.** 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Res* **11:**3–11.
7. **Mardis ER.** 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9:**387–402.
8. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13:**341.
9. **Schneider GF, Dekker C.** 2012. DNA sequencing with nanopores. *Nat Biotechnol* **30:**326–328.
10. **Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD.** 2006. Amplification of complex gene libraries by emulsion PCR. *Nat Methods* **3:**545–550.
11. **Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV.** 2009. The challenges of sequencing by synthesis. *Nat Biotechnol* **27:**1013–1023.
12. **Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM.** 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:**1728–1732.
13. **Arnaud CH.** 2013. DNA sequencing: zero-mode waveguides turn 10. *Chem Eng News* **91:**34
14. **McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A.** 2010. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Lett* **10:**2237–2244.
15. **Maitra RD, Kim J, Dunbar WB.** 2012. Recent advances in nanopore sequencing. *Electrophoresis* **33:**3418–3428.
16. **Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR.** 2015. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* **25:**1750–1756.
17. **Chiu CY.** 2013. Viral pathogen discovery. *Curr Opin Microbiol* **16:**468–478.
18. **Koren S, Phillippy AM.** 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr Opin Microbiol* **23:**110–120.
19. **Quick J, Quinlan AR, Loman NJ.** 2014. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *Gigascience* **3:**22.
20. **Rousseaux S, Khochbin S (ed).** 2011. *Epigenetics and Human Reproduction*, p. 1. Springer, Berlin, Germany.
21. **Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, Bouquet J, Greninger AL, Luk KC, Enge B, Wadford DA, Messenger SL, Genrich GL, Pellegrino K, Grard G, Leroy E, Schneider BS, Fair JN, Martínez MA, Isa P, Crump JA, DeRisi JL, Sittler T, Hackett J Jr, Miller S, Chiu CY.** 2014. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res* **24:**1180–1192.
22. **Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW.** 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLOS Comput Biol* **9:** e1003031.
23. **Margulies M, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:**376–380.
24. **Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW.** 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res* **38:** e159.
25. **Chaisson MJ, Brinza D, Pevzner PA.** 2009. *De novo* fragment assembly with short mate-paired reads: does the read length matter? *Genome Res* **19:**336–346.
26. **Hedegaard J, Thorsen K, Lund MK, Hein AM, Hamilton-Dutoit SJ, Vang S, Nordentoft I, Birkenkamp-Demtröder K, Kruhøffer M, Hager H, Knudsen B, Andersen CL, Sørensen KD, Pedersen JS, Ørntoft TF, Dyrskjøt L.** 2014. Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One* **9:** e98187.
27. **Chai V, Vassilakos A, Lee Y, Wright JA, Young AH.** 2005. Optimization of the PAXgene blood RNA extraction system for gene expression analysis of clinical samples. *J Clin Lab Anal* **19:**182–188.
28. **Wylie KM, Weinstock GM, Storch GA.** 2012. Emerging view of the human virome. *Transl Res* **160:**283–290.
29. **Petti CA.** 2007. Detection and identification of microorganisms by gene amplification and sequencing. *Clin Infect Dis* **44:**1108–1114.
30. **Zwirglmaier K, Ludwig W, Schleifer KH.** 2004. Recognition of individual genes in a single bacterial cell by fluorescence in situ hybridization—RING-FISH. *Mol Microbiol* **51:** 89–96.
31. **Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK.** 2014. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog* **10:**e1004437.
32. **Towner JS, Rollin PE, Bausch DG, Sanchez A, Crary SM, Vincent M, Lee WF, Spiropoulou CF, Ksiazek TG, Lukwiya M, Kaducu F, Downing R, Nichol ST.** 2004. Rapid diagnosis of Ebola hemorrhagic fever by reverse transcription-PCR in an outbreak setting and assessment of patient viral load as a predictor of outcome. *J Virol* **78:**4330–4341.

33. **Kohl C, Brinkmann A, Dabrowski PW, Radonić A, Nitsche A, Kurth A.** 2015. Protocol for metagenomic virus detection in clinical specimens. *Emerg Infect Dis* **21:**48–57.
34. **Marston DA, McElhinney LM, Ellis RJ, Horton DL, Wise EL, Leech SL, David D, de Lamballerie X, Fooks AR.** 2013. Next generation sequencing of viral RNA genomes. *BMC Genomics* **14:**444.
35. **Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ.** 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7:**111–118.
36. **Bystrykh LV.** 2012. Generalized DNA barcode design based on Hamming codes. *PLoS One* **7:**e36852.
37. **Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J Jr, Delwart EL, Chiu CY.** 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* **87:**11966–11977.
38. **Motley ST, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW.** 2014. Improved multiple displacement amplification (iMDA) and ultraclean reagents. *BMC Genomics* **15:**443.
39. **Laurence M, Hatzis C, Brash DE.** 2014. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* **9:**e97876.
40. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW.** 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12:**87.
41. **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A.** 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28:**1647–1649.
42. **Blankenberg D, Hillman-Jackson J.** 2014. Analysis of next-generation sequencing data using Galaxy. *Methods Mol Biol* **1150:**21–43.
43. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:**621–628.
44. **Wang Z, Gerstein M, Snyder M.** 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:**57–63.
45. **Weber G, Shendure J, Tanenbaum DM, Church GM, Meyerson M.** 2002. Identification of foreign gene sequences by transcript filtering against the human genome. *Nat Genet* **30:**141–142.
46. **Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM.** 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**(D1)**:**D633–D642.
47. **Miller JR, Koren S, Sutton G.** 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95:**315–327.
48. **Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu CY, Delwart EL.** 2015. An ensemble strategy that significantly improves *de novo* assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic Acids Res* **43:**e46.
49. **Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL.** 2014. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One* **9:**e105067.
50. **Huson DH, Auch AF, Qi J, Schuster SC.** 2007. MEGAN analysis of metagenomic data. *Genome Res* **17:**377–386.
51. **Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C.** 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9:**811–814.
52. **Wood DE, Salzberg SL.** 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15:**R46.
53. **Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM.** 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* **30:**1033–1036.
54. **Chakravorty S, Helb D, Burday M, Connell N, Alland D.** 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods* **69:**330–339.
55. **Salipante SJ, Sengupta DJ, Rosenthal C, Costa G, Spangler J, Sims EH, Jacobs MA, Miller SI, Hoogestraat DR, Cookson BT, McCoy C, Matsen FA, Shendure J, Lee CC, Harkins TT, Hoffman NG.** 2013. Rapid 16S rRNA next-generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS One* **8:**e65226.
56. **Huson DH, Mitra S.** 2012. Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods Mol Biol* **856:**415–429.
57. **Gürtler V, Stanisich VA.** 1996. New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. *Microbiology* **142:**3–16.
58. **Liu D.**2011. *Molecular Detection of Human Fungal Pathogens*. CRC Press, Boca Raton, FL.
59. **Grice EA, Segre JA.** 2011. The skin microbiome. *Nat Rev Microbiol* **9:**244–253.
60. **Cho I, Blaser MJ.** 2012. The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13:**260–270.
61. **Huttenhower C, et al, Human Microbiome Project Consortium.** 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486:**207–214.
62. **Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG, Knight R.**2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protocols Microbiol* **Chapter 10:**Unit 10.7.
63. **Ahn SH, Tsalik EL, Cyr DD, Zhang Y, van Velkinburgh JC, Langley RJ, Glickman SW, Cairns CB, Zaas AK, Rivers EP, Otero RM, Veldman T, Kingsmore SF, Lucas J, Woods CW, Ginsburg GS, Fowler VG Jr.** 2013. Gene expression-based classifiers identify *Staphylococcus aureus* infection in mice and humans. *PLoS One* **8:**e48979.
64. **Anderson ST, Kaforou M, Brent AJ, Wright VJ, Banwell CM, Chagaluka G, Crampin AC, Dockrell HM, French N, Hamilton MS, Hibberd ML, Kern F, Langford PR, Ling L, Mlotha R, Ottenhoff TH, Pienaar S, Pillay V, Scott JA, Twahir H, Wilkinson RJ, Coin LJ, Heyderman RS, Levin M, Eley B, ILULU Consortium, KIDS TB Study Group.** 2014. Diagnosis of childhood tuberculosis and host RNA expression in Africa. *N Engl J Med* **370:**1712–1723.
65. **Woods CW, McClain MT, Chen M, Zaas AK, Nicholson BP, Varkey J, Veldman T, Kingsmore SF, Huang Y, Lambkin-Williams R, Gilbert AG, Hero AO III, Ramsburg E, Glickman S, Lucas JE, Carin L, Ginsburg GS.** 2013. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. *PLoS One* **8:**e52198.
66. **Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y.** 2013. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One* **8:**e71462.
67. **Mejias A, Ramilo O.** 2014. Transcriptional profiling in infectious diseases: ready for prime time? *J Infect* **68**(Suppl 1)**:** S94–S99.
68. **Croucher NJ, Thomson NR.** 2010. Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol* **13:**619–624.
69. **Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, Sokolic R, Garabedian E, Candotti F, Buckley RH, Reed KD, Meyer TL, Seroogy CM, Galloway R, Henderson SL, Gern JE, DeRisi JL, Chiu CY.** 2014. Actionable diagnosis of

neuroleptospirosis by next-generation sequencing. *N Engl J Med* **370:**2408–2417.
70. **Burd EM.** 2010. Validation of laboratory-developed molecular assays for infectious diseases. *Clin Microbiol Rev* **23:**550–576.
71. **Weiss RL.** 2012. The long and winding regulatory road for laboratory-developed tests. *Am J Clin Pathol* **138:**20–26.
72. **Tripp HJ, Hewson I, Boyarsky S, Stuart JM, Zehr JP.** 2011. Misannotations of rRNA can now generate 90% false positive protein matches in metatranscriptomic studies. *Nucleic Acids Res* **39:**8792–8802.
73. **Zook JM, Samarov D, McDaniel J, Sen SK, Salit M.** 2012. Synthetic spike-in standards improve run-specific systematic error analysis for DNA and RNA sequencing. *PLoS One* **7:** e41356.
74. **Schtig H.** 2014. Development of FDA MicroDB: A Regulatory-Grade Microbial Reference Database. http://www.slideshare.net/NathanOlson/sichtig-h-tallonlmicrodbnist standards.
75. **Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, Stryke D, Bouquet J, Somasekar S, Linnen JM, Dodd R, Mulembakani P, Schneider BS, Muyembe-Tamfum J-J, Stramer SL, Chiu CY.** 2015. Rapid metagenomics identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med* **7:**99.