CURATED COLLECTION

Advances in Molecular Epidemiology of Infectious Diseases

Microbiology Spectrum
American Society for Microbiology Press

Platforms and Analytical Tools Used in Nucleic Acid Sequence-Based Microbial Genotyping Procedures*

DUNCAN MACCANNELL[1]

[1]Office of Advanced Molecular Detection, National Center for Zoonotic and Emerging Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333

ABSTRACT In the decade and a half since the introduction of next-generation sequencing (NGS), the technical feasibility, cost, and overall utility of sequencing have changed dramatically, including applications for infectious disease epidemiology. Massively parallel sequencing technologies have decreased the cost of sequencing by more than 6 orders or magnitude over this time, with a corresponding increase in data generation and complexity. This review provides an overview of the basic principles, chemistry, and operational mechanics of current sequencing technologies, including both conventional Sanger and NGS approaches. As the generation of large amounts of sequence data becomes increasingly routine, the role of bioinformatics in data analysis and reporting becomes all the more critical, and the successful deployment of NGS in public health settings requires careful consideration of changing information technology, bioinformatics, workforce, and regulatory requirements. While there remain important challenges to the sustainable implementation of NGS in public health, in terms of both laboratory and bioinformatics capacity, the impact of these technologies on infectious disease surveillance and outbreak investigations has been nothing short of revolutionary. Understanding the important role that NGS plays in modern public health laboratory practice is critical, as is the need to ensure appropriate workforce, infrastructure, facilities, and funding consideration for routine NGS applications, future innovation, and rapidly scaling NGS-based infectious disease surveillance and outbreak response activities. *This article is part of a curated collection.

INTRODUCTION

The introduction of next-generation sequencing (NGS) has had an important and lasting impact on clinical and public health microbiology. These new technologies hold tremendous promise in improving the speed, accuracy, and resolution of infectious disease diagnostics and the quality and timeliness of laboratory data for public health surveillance and outbreak detection and response. Sustainable implementation of these technologies will require ongoing research and commitment to laboratory and informatics capacity, workforce development, and the development of standardized methods and protocols.

This review focuses on the technological basis for most commercially available NGS platforms, the practical bioinformatic and technical requirements that are necessary for implementation, and a number of the key technical and scientific challenges yet to be resolved.

Received: 25 July 2018, **Accepted:** 28 November 2018, **Published:** 8 February 2019

Editors: Lee W. Riley, Divisions of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, Berkeley, CA; Ronald E. Blanton, Center for Global Health & Diseases, Case Western Reserve University, Cleveland, OH

Citation: Maccannell D. 2019. Platforms and analytical tools used in nucleic acid sequence-based microbial genotyping procedures. *Microbiol Spectrum* 7(1):AME-0005-2018. doi:10.1128/microbiolspec.AME-0005-2018.

Correspondence: Duncan MacCannell, fms2@cdc.gov

Curated Collection: Advances in Molecular Epidemiology of Infectious Diseases.

FIRST-GENERATION DNA SEQUENCING PLATFORMS

In the late 1970s, several different DNA sequencing methods were introduced, establishing the first generation of widely accessible technologies for genetic sequencing. The first of these techniques, introduced by Allan Maxam and Walter Gilbert, relied on chemical modification and base-specific cleavage of DNA molecules, whereas the second approach, proposed by Frederick Sanger and colleagues, relied on selective chain termination due to the incorporation of dideoxyribobases into a growing DNA strand. Of the two methods, Maxam-Gilbert sequencing was initially the more popular choice for most laboratories, since early iterations of Sanger sequencing required extensive cloning and sample preparation, while the Maxam-Gilbert method could be performed more or less directly on any purified DNA sample. With the development of PCR in 1983, however, Sanger sequencing shed many of its initial limitations and rapidly outpaced Maxam-Gilbert, as subsequent refinements made it by far the more feasible and cost-effective choice. By the middle to late 1990s, multichannel capillary sequencers began to replace traditional polyacrylamide slab gels for the separation of DNA fragments and the interpretation of the corresponding sequence. Today, capillary gel electrophoresis systems typically have between 8 and 24 capillary channels and can perform a range of different fragment analysis tasks on samples arranged in 96- or 384-well plates.

The development of capillary Sanger sequencing was a critical technological advance that enabled early, large-scale sequencing projects like the Human Genome Project. Sanger sequencing remains a fixture of many molecular biology laboratories even today, despite the advent of NGS technologies, where it serves an important role, particularly for target validation, specialized or small-scale sequencing applications, and high-throughput molecular surveillance where NGS is not yet feasible or cost competitive.

Sanger Sequencing: Mechanism

In order to understand the molecular basis of Sanger sequencing, it helps to think of the chain termination reaction as simply a derivative of DNA amplification strategies, such as PCR. Like these methods, Sanger sequencing requires a template DNA molecule to sequence, a high-fidelity DNA polymerase, a sequencing primer to target and prime the reaction, deoxyribonucleotides (dATP, dTTP, dGTP, and dCTP) to extend the chain, and also a much smaller proportion of fluorescently labeled dideoxyribonucleotides (ddATP, ddTTP, ddGTP, and ddCTP) that are incorporated randomly into the chain and terminate any further elongation. Dideoxynucleotides (ddNTPs) are typically added to the reaction mixture in concentrations that are 100- to 1,000-fold lower than their corresponding deoxynucleotide (dNTP) equivalents. As the DNA polymerase extends the nascent chain, it may incorporate either a dNTP or a ddNTP at any given base position. Should a ddNTP be incorporated, that particular molecule becomes terminated, since the ddNTP lacks the necessary 3′ hydroxyl (OH) group for the next polymerization step. The ddNTP moiety typically includes a specific fluorophore, so that both the fragment size and the identity of the 3′ terminal base can be identified by the sequencing instrument or its operator.

As the sequencing reaction proceeds, ddNTPs will gradually incorporate into and terminate the polymerization reaction at each and every base position along the chain. At the completion of the sequencing reaction, each reaction tube will ultimately contain a set of labeled, single-stranded DNA fragments that represent all of the possible termination sites for each of the four nucleotide bases (A, T, G, and C). By running this mixture of fragments on a high-resolution capillary sequencer, it is possible to separate and resolve these fragments with base-level accuracy and thereby determine the sequence of the original template strand.

Earlier iterations of Sanger sequencing used P32 radiolabeling in the place of individual fluorophores and required laboratorians set up four separate reactions for A, T, G, and C chain termination. These four reactions were then run as separate lanes on a high-resolution polyacrylamide sequencing gel, and the operator would then read the corresponding sequence from 5′ to 3′, starting with the largest of fragments. Fluorescent labeling greatly improved the efficiency, safety, and automation of Sanger sequencing by allowing all four reactions to be run in the same reaction tube, with automated software calling bases directly from the instrument detector array.

Public Health and Clinical Applications

Capillary Sanger sequencing is an important and established cornerstone of both diagnostic testing and molecular surveillance and is widely used in many clinical and public health microbiology laboratories. Even as NGS platforms begin to see more widespread adoption, Sanger sequencing offers distinct advantages over NGS in terms of sequence length (>800 bp), sequence quality, regulatory familiarity, and cost per sample. These are critical advantages for assay development and

validation and are particularly important for large-scale molecular surveillance applications and high-throughput diagnostics, where NGS remains impractical and expensive or it is difficult to batch samples efficiently.

For example, Sanger sequencing is still the preferred approach to characterize many viral infections, including HIV, for which it is used for both strain typing (partial sequencing of *pol* hypervariable regions) and the prediction of antiviral susceptibility (1). Sanger sequencing also remains a preferred approach for the identification and categorization of many bacterial antimicrobial resistance markers, including acquired carbapenemase resistance genes among *Enterobacteriaceae*, confirmation of clinically important staphylococcal cassette chromosome *mec* element and leucocidin variants among methicillin-resistant *Staphylococcus aureus* and the reference identification of bacterial, fungal, and protozoan pathogens using 16S/18S/internal transcribed spacers (2–5). Multilocus sequence typing (MLST) and even some single-locus sequencing approaches, such as streptococcal *emm* typing, remain important applications for molecular epidemiology and public health (6). In many public health laboratories, capillary sequencers are also used for other, nonsequencing fragment analysis tasks, including multilocus variable number tandem repeat analysis for strain typing enteric bacteria and high-resolution PCR ribotyping for *Clostridium difficile* (7, 8). Given the flexibility of capillary gel electrophoresis systems, some enduring cost and throughput advantages, and the multitude of workhorse diagnostic assays that are used in laboratories throughout the world, Sanger sequencing will likely continue to play an important role in surveillance and diagnostics for the foreseeable future, despite recent advances in NGS technology.

NEXT-GENERATION SEQUENCING

The first commercially available next generation sequencing instruments revolutionized the biological sciences when they were first introduced in 2005. This introduction resulted in a near-immediate and dramatic decrease in the overall cost of sequencing and a corresponding increase in the volume and complexity of sequence data produced. These massively parallel, short-read sequencers generated read lengths of several dozen to several hundred base pairs and were based on breakthrough discoveries during the 1990s and years of intense research and development from a number of different academic and commercial groups. In general, these systems had significant advantages over conventional dye terminator Sanger sequencing and could be applied to a wide range of microbiologic assays and genomic studies for basic science, clinical, and public health research. In the decade that has followed, NGS instrumentation and its surrounding technology space have seen extremely rapid evolution, with new applications and platforms and iterative improvements to existing sequencing instruments, reagents, and consumables, kits for sample preparation and processing, and support instrumentation for NGS workflows. Today, a number of different NGS platforms and technologies are available, with a range of different form factors, performance characteristics, error models, and operational requirements. Although this chapter is not meant to be a definitive historical overview of NGS systems, the characteristics of many of the most of the common NGS systems are summarized in Table 1 (9).

MICROBIAL NEXT-GENERATION SEQUENCING LABORATORY WORKFLOWS

For clinical and public health microbiology, NGS represents a powerful and flexible tool for the study of infectious diseases. NGS workflows can accommodate DNA and/or RNA from a variety of different sources, including host, vector, pathogen, and the environment, and may include whole sequences, fragments, or selectively amplified targets. In general, NGS workflows consist of five or six principal steps, beginning with an enriched specimen or pure culture isolate, and include (i) DNA/RNA extraction, isolation, conversion, quantitation, and quality assessment; (ii) enzymatic or mechanical shearing, size selection, and NGS library construction; (iii) presequencing cleanup and quality assurance; (iv) sequencing; and, finally, (v) downstream data processing, bioinformatic analysis, and result reporting (Fig. 1). The relative consistency of NGS laboratory protocols presents some important opportunities for technical and resource consolidation, and efforts are currently under way to develop and validate standardized sequencing workflows that can support a range of different pathogens and sequencing applications.

DNA/RNA Extraction

Depending on the pathogen and the type of sequencing that is being performed, DNA and RNA extraction may use vendor-supplied kits or rely on classic methods, such as phenol-chloroform extraction. The type of microorganism, the nature and quality of the specimen, and both the context and purpose of the assay are all critical factors in maximizing the quality and yield of sequenceable material. Applications that require the sequencing of

TABLE 1 Summary characteristics of most common NGS commercial systems[a]

Company or product type	Instrument name	Form factor	Sequencing technology[c]	Instrument cost	Read length (typical)	Runtime (modes)	Single-read output (modes)	Maximum sequence output	Sequence cost (reagents; USD/Mb)	Single-pass error rate (type[d])
Illumina	iSeq 100	Benchtop/ portable	SBS / CMOS	$19K	2 × 150 bp	9 h/18 h	25M	1.2 Gb	$0.50	0.1% (a)
	MiniSeq	Benchtop	SBS	$50K	2 × 150 bp	24 h/17 h	25M/8M	7.5 Gb	$0.20	0.1% (a)
	MiSeq(Dx)	Benchtop	SBS	$99K	2 × 250 bp	56 h	25M	15 Gb	$0.10	0.1% (a)
	NextSeq(Dx)	Benchtop	SBS	$250K	2 × 150 bp	29 h/26 h	400M/130M	120 Gb	$0.03	0.1% (a)
	HiSeq	Capital	SBS	$750K	2 × 150 bp	6 days/40 h	2B/2B	1 Tb	$0.03	0.1% (a)
	NovaSeq	Capital	SBS	$850K	2 × 150 bp	16 h/44 h	20B	6 Tb	$0.01	0.1% (a)
Thermo Fisher	PGM[b]	Benchtop	Semiconductor	$50K	400 bp	7 h	5.5M	2 Gb	$0.40	~2% (b, c)
Ion Torrent	Proton	Benchtop	Semiconductor	$150K	200 bp	4 h	83M	10 Gb	$0.06	~2% (b, c)
	S5	Benchtop	Semiconductor	$65K	400 bp	2.5 h/4 h	5M/80M	15 Gb	$0.08	~2% (b, c)
Pacific	RSII	Capital	SMRT	$700K	10,000–15,000 bp	4 h	50K	1 Gb	$0.30	10–15% (b)
Biosciences	Sequel	Capital	SMRT	$350K	10,000–20,000 bp	6 h	350K	7 Gb	$0.18	10–15% (b)
Oxford Nanopore	MinION MkIb	Portable	Nanopore	$1K	>10,000 bp	1 min to 48 h	2.2M/4.4M	Up to 40 Gb	$0.05–$0.10	5–15% (b, c)
	PromethION	Benchtop	Nanopore	$160K	>10,000 bp	1 min to 48 h	625M/1.25B	15 Tb	<$0.01	5–15% (b, c)
Roche 454	FLX+[b]	Benchtop	Pyrosequencing	$100K	650 bp	20 h	1M	650 Mb	$10.00	1% (b, c)
Applied Biosystems	3730xl	Capital	Sanger	$100K	700 bp	2 h	96	67 kb	$2,800.00	0.1% (a)

[a]References: http://www.illumina.com; http://www.pacb.com; https://www.thermofisher.com; https://www.nanoporetech.com. K, thousand; M, million; B, billion.
[b]Included for reference; deprecated.
[c]SBS, sequencing by synthesis; SMRT, single molecule real time.
[d]Error type: substitution (a), indel (b), or homopolymer (c).

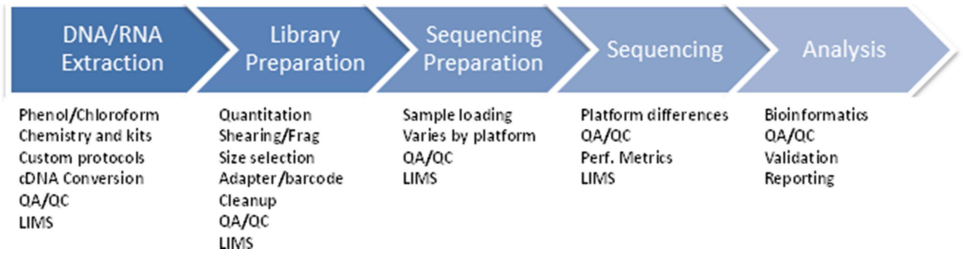

FIGURE 1 Principal steps of a generic NGS workflow.

RNA or total nucleic acid generally include a cDNA conversion step, although direct RNA sequencing is becoming more and more feasible, and this too may also have important effects on quality and yield. Another important consideration, regardless of the specific extraction protocol that is being used, is that all of these methods introduce some level of bias due to preferential lysis, recovery loss, and other factors that may be applied unevenly across a specimen or sample. For this reason, it is important to validate extraction protocols carefully, to select the most appropriate method, and to ensure that the impact on qualitative and quantitative sequencing projects is well understood.

Library Construction

Library construction is required for all commercial NGS platforms and typically involves a number of discrete steps. First, the template DNA is sheared or broken into smaller fragments; fragments are then sorted and selected by size, NGS adapter and barcode sequences are then attached, and the library is subjected to a final quality assessment and cleanup before sequencing. Template DNA can be sheared to sequenceable lengths using either mechanical or enzymatic approaches. Mechanical shearing is usually performed using pulsed ultrasonic disruption of the DNA using specialized instrumentation. Enzymatic shearing, on the other hand, is usually kit based and can use either endonuclease- or transposon-based methods.

Library construction is extremely time and labor-intensive, particularly at scale. For this reason, appropriate laboratory automation (e.g., robotics), laboratory information management systems, and sample management strategies are critical. Another important consideration is the overall cost and sophistication that are required to sequence at scale. While large sequencing centers typically use high-throughput, focused ultrasonication to shear genomic DNA into appropriate fragment lengths for sequencing, the cost of this equipment is often a barrier to lower-volume laboratories or resource-limited settings. Enzymatic library construction methods, such as Illumina Nextera, which relies on random transposase-mediated fragmentation and adapter insertion, have been shown to produce analogous results for most sequencing applications and are generally more feasible for widespread use. For this reason, public health activities that include large-scale distributed sequencing, (for example, PulseNet [http://www.cdc.gov/pulsenet]), have generally standardized on enzymatic library construction and have carefully validated and weighed the inclusion of additional protocol steps to limit the need for expensive or complicated ancillary equipment. While the protocols for library construction are generally platform specific, and libraries cannot be easily interchanged between platforms, many of the steps in the overall library construction process remain fundamentally similar in terms of their conditions, biases, and limitations.

An increasing number of laboratories have begun to include NGS-based microbial analyses under quality-managed or Clinical Laboratory Improvements Amendment (CLIA)-compliant frameworks. Implementing NGS-based methods for routine use under CLIA or other quality management system frameworks is a complex undertaking relative to conventional molecular testing. Quality managers must consider the integrity and reproducibility of the entire analytical process, from raw NGS sequence data to the final reported results, and provide a verifiable audit trail for all bioinformatic steps, software, and databases used in the analysis. Laboratories also must consider how to implement meaningful quality control and quality assurance measures, set appropriate validation criteria, and establish realistic competencies and proficiency testing for individual bioinformatic analyses across the entire NGS workflow (10).

Different commercially available NGS platforms are discussed in the following section.

NGS PLATFORMS AND INSTRUMENTATION

Short-Read Sequencing Technologies

Short-read NGS sequencing instruments perform highly parallelized sequencing of fragmented DNA, outputting

vast numbers of short sequence reads that range in size from 75 to 800 bp in length, depending on the hardware platform and sequencing configuration required. Because of the versatility, massive output, and relatively low cost of ownership and operation, short-read sequencing has rapidly become an established workhorse technology in many academic, commercial, clinical and public health laboratories throughout the world. Despite these advantages, short-read sequencing can present important challenges for downstream bioinformatic analysis and interpretation in many instances, particularly in the sequencing of genomes with complex structure or extensive repeats or in specific applications where mixed assembly, recombination, or phasing may be of concern. In these instances, short-read sequencing data are often paired with long-read sequencing for hybrid assembly, optical mapping, or validated using orthogonal methods.

The current market for short-read sequencers covers a range of different form factors and sequencing capacities, and instrument selection will depend on a given laboratory's anticipated throughput and sequencing needs and the amount of local infrastructure and support available, including laboratory, personnel, and information technology resources. Compact benchtop sequencers, such as the Illumina MiSeq and Thermo Fisher Ion Torrent, cost in the range of $65,000 to $100,000 U.S. dollars (USD) and are becoming increasingly standard equipment in most low-volume academic, clinical, and public health research laboratories. Smaller benchtop instruments, such as the MiSeq, are equally at home in high-volume core sequencing laboratories, where they are often used for small-scale or time-sensitive sequencing tasks or to assess the quality of libraries before they are run on a larger, more costly instrument, such as the Illumina HiSeq. For its part, the Illumina HiSeq, which comes in a range of models and configurations, is usually the preferred choice for industrial and high-throughput sequencing laboratories. For smaller laboratories, or those with variable sample volumes, the cost and complexity of these instruments limit the feasibility of on-premise use, and such laboratories tend to favor midrange sequencing platforms and the outsourcing of larger tasks to academic core labs or commercial sequencing facilities (9, 11).

There are important differences between the two vendors in terms of instrumentation and target markets. In recent years, for example, Illumina has focused primarily on general-purpose, short-read sequencing across a range of different applications, while Thermo Fisher has largely shifted their product development efforts to highly multiplexed amplicon sequencing for oncology, infectious diseases, and other clinical markets, even going so far as to adapt their AmpliSeq massively parallel amplicon sequencing chemistry to the Illumina platform. The technological basis, strengths, and weaknesses of both the Ion Torrent semiconductor sequencing and Illumina reversible dye terminator sequencing systems and platforms are discussed in the section that follows.

Sequencing by synthesis (SBS; Illumina)

Illumina short-read sequencers are currently a predominant platform in many NGS markets, with a range of instruments based on a proprietary reversible dye terminator approach. In general, sequencing libraries for Illumina instruments are interchangeable, meaning that a sample prepared for sequencing on a MiSeq may be run for similar purposes on a larger instrument, such as the HiSeq 2500 or 4000.

There are a number of different ways to prepare DNA for Illumina sequencing, but most of them involve fragmenting the template sequence into inserts between 150 and 500 bp in length and attaching specific adapters to each end of the sequence that include sequence binding site, one or more barcode indices, and terminal sequences that are complimentary to the two flow cell oligonucleotides. Once the sequencing libraries have been prepared, quantified, and assessed for quality, they are ready to be loaded. The instrument itself goes through two principal phases over the course of a sequencing run: (i) cluster generation, in which bridge amplification is used to focally amplify clusters of each insert sequence that has been seeded randomly across the flow cell surface, and (ii) sequencing, in which the sequence of each of these clusters is determined by stepwise synthesis of the complement strand.

Cluster generation

Inside each flow cell lane, a dense lawn of two different oligonucleotide probes is covalently attached to the glass surface. When single-stranded sequencing libraries are applied to the flow cell lanes, the complementary adapter sequences on each fragment hybridize to one of the two different oligonucleotides that make up the lawn. A high-fidelity DNA polymerase is then added to generate the complement strand to the template, the resulting double-stranded template is denatured, and the original template fragment is washed away. The other end of each template has the corresponding adapter, which bears the other oligonucleotide complement. Amplification of the template occurs through an isothermal process known as bridge amplification: the tethered, single-stranded template sequence bends, and the other

end hybridizes to a nearby, covalently attached probe, forming a bridge. A polymerization step makes a second copy of the template sequence, which is then denatured to repeat the amplification process. This process occurs all over the surface of the flow cell, amplifying all of the fragments that were seeded to each lane, and results in the formation of hundreds of millions of individual clusters. Once bridge amplification is complete, the reverse strands are all cleaved and washed away, leaving the clusters of amplified template correctly oriented and ready to sequence.

Sequencing by synthesis

Before sequencing begins, the 3′ end of the fragment is blocked to prevent mispriming and the first sequencing primer is added to the flow cell, where it binds to the sequencing binding site in the 3′ adapter region of each strand. For each cycle of sequencing, fluorescently labeled dNTPs are added to the flow cell, and the primer is extended by a single base. At the end of each sequencing cycle, the instrument laser excites the flow cell at a specific wavelength and images the clusters in either 2 or 4 different colors, depending on the instrument and runtime configuration. Using high-resolution image processing, the instrument is able to determine the base that was incorporated by each cluster at any given cycle using the observed fluorescence signature. The dye terminator is then cleaved from the 3′ end of the nascent strand and the next cycle of extension and imaging begins. Once the sequence read is complete, the strand is denatured, and the sequence product is washed away. An indexing primer is then added and used to sequence the first index for each cluster in the 5′ adapter sequence of the template strand.

Illumina read lengths can be anywhere between 50 and 300 bases in length, and sequencing is frequently carried out using a paired-read approach, meaning that the system generates a pair of sequence reads that fall within a certain known distance of each other on the template. In order to sequence the other side of the sequence pair, the 3′ end of the fragment is deprotected and allowed to hybridize once more with a nearby covalently attached probe sequence. During a single round of bridge amplification, the second index sequence in the 3′ end of the original template is read, and the complement of the original template strand is generated. Once complete, the 3′ end of the nascent strand is blocked, the bridge is denatured, and the original template strand is cleaved and washed away, leaving only the blocked complement strand. Sequencing for the second read in the pair proceeds exactly as described above using the second sequencing primer.

Two-color versus four-color SBS

Because of differences in the underlying detector technology—the NextSeq and MiniSeq both use two-color SBS detection, whereas the MiSeq and HiSeq typically use four—the newer NextSeq and MiniSeq instruments achieve significantly higher sequencing speed and output at a lower cost, with a relatively marginal decrease in base calling accuracy. The decrease in accuracy was initially much more severe, and the first system to implement 2-color SBS, the NextSeq, was often limited to resequencing tasks as a result. Subsequent improvements to the base calling algorithms, sequencing chemistry, and consumables have greatly improved the accuracy of two-color SBS for both the NextSeq and MiniSeq platforms, and both instruments are now commonly used interchangeably with four-color SBS instruments for a range of sequencing tasks, including *de novo* sequencing and assembly. The company recently introduced the iSeq 100 as a low-cost, complementary metal oxide semiconductor (CMOS)-based benchtop instrument; the iSeq streamlines SBS even further, using a one-color approach with two chemistry and two imaging steps to identify bases (12).

Cluster density

Shortly after sequencing begins—during the first 20 cycles or so—the instrument scans the flow cell to determine cluster density of each lane. This is an important runtime metric of the overall loading efficiency of each lane, and it impacts both the output and quality of sequence data from the run. On a MiSeq, for example, cluster densities of 700,000 or 1 million clusters per square millimeter are fairly optimal, depending on the specific sequencing chemistry and loading parameters that were used. Cluster density is also an important metric for quality assurance/quality control and a useful early indicator of protocol or procedural issues.

Semiconductor sequencing (Ion Torrent)

Semiconductor sequencing is a type of sequencing by synthesis that relies on the detection of hydrogen ions that are released during DNA polymerization using highly tuned ion-sensitive field effect transistor CMOS detection. The underlying technology was originally developed by a research team at Imperial College London, who formed a spin-off company called DNAe (DNA Electronics) to develop and license the intellectual property. While the most notable implementation of semiconductor sequencing is the Thermo Fisher Ion Torrent line of sequencers, the technology has also been licensed to Roche 454, GenapSys, and others for both sequencing

and direct macromolecular detection, and it remains under active development (13).

Sample preparation and library construction for Ion Torrent sequencing are relatively standard, and most nucleic acid extraction, quantification, quality control, and fragmentation protocols for NGS may be adapted to the Ion Torrent sequencing workflow. Once the template sequence has been appropriately fragmented and assessed for quality and size, Ion Torrent-specific adapter sequences are attached to each fragment in the library by ligation or PCR. These adapter sequences are used to anneal the library fragments to the surface of microbeads, which are then clonally amplified using emulsion PCR (emPCR). In general, the reaction stoichiometry is such that a single library fragment is amplified across the surface of each individual microbead; however, a small proportion of the beads may include multiple fragments and result in mixed or polyclonal amplification. Following amplification, the microbeads are deposited onto the Ion Torrent chip using a brief centrifugation step, such that each well on the chip will contain a single template-coated microbead.

As sequencing is initiated and the run proceeds, the reaction surface of the chip is sequentially flooded with each unmodified dNTP (dATP, dTTP, dGTP, and dCTP) in a repeating cycle. The nucleotide that is released during each step of the cycle will be incorporated only in those microwells where the sequence is complementary to the underlying template. When this polymerization occurs, it releases a hydrogen ion and pyrophosphate (PP_i) as byproducts, and the resulting minute change in pH is detected by an ion-sensitive field effect transistor-based detector in the base of each microwell on the chip. If the polymerase encounters homopolymer sequences in the template strand, multiple dNTPs may be incorporated during a single instrument step. In this instance, the instrument is calibrated to determine the number of homopolymer bases that were incorporated based on relative changes in the local pH within each microwell (14).

There are important theoretical advantages to semiconductor sequencing, namely, in terms of the speed and mechanism of signal detection. Unlike Illumina or Sanger sequencing, for example, where base calling requires laser excitation and detection of the corresponding fluorescence, semiconductor sequencing can detect microscale changes in pH and assign base calls without the need for complicated optics or image processing algorithms. This results in much faster sequence throughput and reduces the complexity of the instrument path required for base detection to the point where it can be miniaturized and parallelized in each well across the chip.

Conversely, because of the minute scale of the pH changes that are measured by the instrument at each cycle, semiconductor sequencing is also exquisitely sensitive to changes in ionic concentration and conditions that impact the surface chemistry of the chip. As a result, the first several generations of Ion Torrent instruments (e.g., the PGM and the Proton) required sensitive calibration and the use of extremely pure molecular-grade water and argon carrier gas for sequencing operations. Relative to competing platforms, such as the Illumina MiSeq, the stringency of these requirements introduced important challenges for many laboratories that were looking to implement NGS, particularly those with logistical constraints or in resource-limited settings. Signal sensitivity is also an important source of systematic error. Because of the difficulty associated with the precise quantitation of the microscale pH changes, semiconductor sequencing is prone to homopolymer errors, since the instrument is unable to consistently resolve the exact number of bases incorporated in each well during any given cycle (15).

Ion Torrent is well suited to resequencing and targeted amplicon sequencing, where their AmpliSeq sequencing panels show particular strength. AmpliSeq allows simultaneous amplification and sequencing of up to 6,144 primers across 96 samples on a single Ion 318 chip and has been successfully adapted to support a range of human and microbial applications, including broad-based microbial identification from clinical samples and the identification and sequence-based characterization of acquired antimicrobial resistance genes (14). As a result of this success, Thermo Fisher seems to have increased their emphasis on targeted sequencing panels in recent years, with a corresponding decrease in the prioritization of *de novo* sequencing applications, where other sequencing technologies tend to be more competitively positioned.

Historical short-read sequencing platforms

In the decade and a half since short-read NGS systems were first introduced, several other historically important technological approaches to sequencing have come and gone. While the following three hardware platforms are no longer commercially available, their technological approaches and biases are worth reviewing here, since they are often referred to in the literature and data from these platforms often surfaces in routine searches of the NCBI Sequence Read Archive and other large sequence repositories.

Roche 454 pyrosequencing (deprecated)

When 454 Life Sciences released the GS20 sequencer in 2005, it was the first commercially available next-

generation instrument on the market, and it marked an important and fundamental change in both sequencing and the biological sciences. In the years that followed, iterative improvements to the 454 platform resulted in the FLX, FLX Titanium, and FLX+ Titanium sequencers, with read lengths that approached 650 bp and output that approached 700 million bases. 454 Life Sciences was acquired by Roche in 2007 and closed down in 2013, as the platform was no longer competitive with other NGS platforms and technologies.

454 used a massively parallel pyrosequencing approach to sequence libraries of up to 800 bp in length. Once the template DNA was fragmented, it was repaired, ligated to biotinylated adapter sequences, and titrated onto streptavidin microbeads, such that a single fragment was bound to each individual microbead. emPCR was then used to enclose each bead in a microreaction vessel and to amplify the template fragments across the surface of each bead. The microbeads were then recovered, and the fragments were denatured, leaving uniform single-stranded copies of the template sequence. The suspension of beads was assessed for quality, and a second set of beads with the immobilized DNA polymerase, luciferase, and ATP sulfurylase attached was added. The bead mixture was then applied to the PicoTiterPlate fiber optic slide, allowing the beads to settle into a lattice of 29 microliter wells on the bottom of the channel. A sequencing primer was then bound to the adapter sequence, on each fragment, and sequencing was initiated.

During each cycle, each nucleotide (dATP, dTTP, dGTP, and dCTP) was washed over the plate sequentially. If a base complemented the underlying template strand, it was incorporated by the DNA polymerase, releasing a PP_i during the reaction. The sulfurylase then catalyzed a reaction between PP_i and adenosine 5′-phosphosulfate, which had been added to the solution, to form ATP. In the presence of ATP, luciferase then catalyzed the conversion of luciferin in the solution to oxyluciferin, with a corresponding flash of visible light. By detecting this pulse of light and its intensity, the instrument could then determine the base incorporated at each well. And like semiconductor sequencing, Roche 454 pyrosequencing was also prone to homopolymer errors, since the instrument was often unable to accurately quantify the incorporation of multiple bases during a given cycle.

Helicos tSMS (deprecated)

Helicos BioSciences was established in 2003 and introduced a process that it called true single molecule sequencing (tSMS), with the first instrument to use direct imaging of individual DNA molecules for NGS. Uptake of the Helioscope sequencing platform was relatively modest throughout the early 2000s, and the company filed for bankruptcy in 2012.

The technology behind Helicos sequencing was relatively straightforward. For optimal sequencing, the template DNA was sheared into fragments of 100 to 200 bp in length, and the fragments were assessed for quality. A 50-mer poly(A) tail was then ligated onto the 3′ end of each fragment, with a terminal fluorescent adenosine tag. The fragments were then denatured and applied to the sequencing cell, where a dense lawn of covalently linked poly(T) capture probes hybridized the labeled fragment library to the flow cell surface. Once hybridized, the sequencing flow cell was loaded onto the instrument and the fragments were visualized using laser excitation and high-resolution confocal microscopy to map the location of each individual fragment using fluorescence in a series of images. The fluorescent label was then enzymatically cleaved from the fragment library and washed from the flow cell. The instrument would then initiate the sequencing process by cycling the four labeled nucleotides through the flow cell sequentially and polymerizing a single base extension. At each step, the flow cell was illuminated and reimaged to determine which fragments incorporated the base. The fluorescent label attached to the nucleotides served as a reversible chain terminator that prevented the incorporation of multiple bases per cycle. This label could be cleaved and washed away at the end of each cycle to enable extension for the next nucleotide.

Due to the long time required for base-by-base cycling, illumination, and high-resolution imaging, the runtime on a standard 120-cycle Helioscope run was approximately 8 days, generating over 1 billion short, 32-bp reads and 25 to 35 Gb of raw sequence data. These extremely short fragments posed important bioinformatic assembly, particularly for sequencing large genomes and complex structure. Error was also an important consideration: because the instrument visualized the fluorescence of individual DNA molecules, it was also prone to errors in signal detection and resolution, which also impacted downstream analysis and assembly to a significant extent (16, 17).

ABI SOLiD (deprecated)

Sequencing by oligonucleotide ligation and detection (SOLiD) was another early NGS technology introduced by Applied Biosystems/Life Technologies in 2006. Like other early massively parallel sequencing technologies, the SOLiD platform relied on immobilizing template fragments on magnetic beads, with subsequent colony

amplification by emPCR. Following emPCR, the beads were deposited onto a glass slide and covalently bound to the surface. All fragments were anchored to the bead with a single adapter sequence (P1), which was used to prime the sequencing reaction for all beads across the reaction surface of the slide.

The SOLiD reaction relied on a set of specific 8 base probes, each bearing a 3′ hydroxyl and a 5′ fluorophore with a chemistry known as two-base encoding. For each probe, the first two bases were complementary to the template sequence, while the remaining 6 were degenerate. Once the first two bases had been incorporated, the three terminal bases of the probe were cleaved off, leaving an exposed hydroxyl group on the 5th base of the probe, and fluorescence was measured and cleaved. In the next sequencing cycle, the corresponding probe bound to the 5′ hydroxyl, matching the two subsequent bases. In this way, at the end of the sequencing round, the template strand would have been sequenced with a repeating pattern of two exact base matches, separated by 3 bases of degenerate nonmatching sequence. The sequencing instrument went through up to 5 rounds of sequencing with new universal primers that were each offset by 1 bp. Thus, subsequent sequencing rounds began at positions n-1, n-2, n-3, and n-4, ensuring that the entire template sequence was tiled with exact, two-base matches. In total, the instrument sequenced templates of approximately 25 to 50 bp, generating up to 60 Gb of output sequence. Because of the two-base encoding approach, each base position was read twice, and this internal redundancy pushed the short-read accuracy rate to approximately 99.94%.

Despite its low cost per base, and similar output volumes, SOLiD struggled to compete with Illumina sequencing, largely due to the relatively long runtime of the SOLiD instrument (1 to 2 weeks) and its limited ability to resolve complex palindromic sequences. Although Applied Biosystems still sells systems based on the SOLiD chemistry, high-throughput sequencing on these instruments is generally uncommon in microbial genomics and molecular epidemiologic studies.

Long-Read and Single-Molecule Sequencing

Single-molecule, real-time sequencing (Pacific Biosciences)

Long-read sequencing first rose to common use with the introduction of the Pacific Biosciences RS sequencer in 2010. While early versions of the platform were plagued by relatively high rates of error, the current generation of instruments, sequencing chemistry, and signal processing algorithms have improved the utility and reliability of PacBio long-read sequencing to the point where bacterial genomes are routinely closed as high-quality draft sequences, either with or without accompanying short-read sequence data for hybrid assembly. PacBio instruments perform single-molecule, real-time (SMRT) sequencing and generate hundreds of thousands of reads with average read lengths of 3 to 10 kb. Because the instrument sequences individual DNA molecules, PacBio sequencing has become increasingly useful for deep sequencing and metagenomic applications, and further, because the system can also detect methylated bases by differences in reaction kinetics, epigenetic data are collected simultaneously during the run. Despite these advantages, PacBio sequencing has remained largely impractical for many laboratories due to the size, cost ($750,000), and infrastructure requirements for the large PacBio RSII instruments. In 2016, Pacific Biosciences introduced the Sequel, an instrument jointly developed with Roche. The Sequel has a significantly smaller physical footprint and seems well positioned for both research and routine production sequencing, with significantly higher sequence output (7 to 8 times the output of an RSII), and lower up-front capital costs ($350,000), fewer facility considerations, and lower total cost of ownership.

Unlike for short-read sequencers, which typically use an insert size of less than 500 bases, the average insert size for a PacBio sequencing library is in the range of 3 to 10 kb in length. For most routine PacBio sequencing, a larger insert, of 10 kb, is preferred. For applications where higher base calling accuracy is needed, a smaller (3-kb) insert library can be used for circular consensus sequencing (CCS), in which each base position in the molecule is read at a minimum of three times. Once the template sequence has been sheared into these large fragments, the ends of each fragment are repaired and specialized hairpin adapter sequences are ligated onto the ends of each fragment to create a dumbbell-shaped continuous loop (SMRTbell). Once the adapter sequences have been attached, the sequencing library undergoes a size selection step and is assessed for quality before loading onto the sequencer.

Sequencing on the Pacific Biosciences platform is often referred to as single-molecule, real-time sequencing since the instrument is in fact sequencing individual DNA molecules in real time. In order to load the instrument, the sequencing library is applied to the SMRT cell such that one SMRTbell-adapted fragment is deposited into each of the more than 100,000 zero-mode waveguide (ZMW) microwells on the RSII SMRT cell surface. A ZMW is a physical well structure that is backlit from

below with an excitatory frequency of light and is physically constrained such that the wavelength is unable to penetrate to the top of the well. The dimensions are small—for the RSII SMRT cell, each ZMW is on the order of 20 zeptoliters (20×10^{-21} liters)—creating a reaction volume that illuminates for even the incorporation of a single base on a single strand. During sequencing, a strand-displacing polymerase that is covalently bound to the bottom of the ZMW microwell opens the SMRTbell structure into a circular, single-stranded template and proceeds to sequence the forward and reverse strand of the insert. As dNTPs are incorporated into the strand, their phospholinked fluorescent dye label is cleaved and emits a fluorescent pulse and quickly diffuses out of the ZMW. Sequencing proceeds until the run completes or the polymerase is depleted. Repeated cycles of the template sequence in each ZMW improve the output quality of the sequence: per-base quality scores increases linearly with the number of times each base position is read.

Another important advantage of PacBio SMRT is that it is possible to directly identify certain epigenetic features over the course of normal sequencing operations. Methylation of bases in the template sequence affects the kinetics of the DNA polymerase as it passes over them, and by analyzing differences in the spacing between sequence pulses (the interpulse duration), it is possible to computationally identify certain base modifications, including 6-methyl adenine, 4-methyl cytosine, and Tet-converted 5-methyl cytosine, directly from the sequence data. The bacterial methylome is believed to play an important role in the restriction and modification of bacterial genomes and the gain or loss of genes related to virulence and antimicrobial resistance. A number of studies have begun to leverage PacBio epigenetic data to understand changing patterns in antimicrobial susceptibility (18, 19).

There are also some important challenges with PacBio sequencing, discussed below.

Error rate

The first challenge is the relatively high single-pass error rate. While there has been significant improvement to the chemistry and base calling algorithms in recent iterations, and many of the random errors may be addressed with significant depth or CCS, the error rate of the instrument may still present challenges to certain sequencing projects.

Sample requirements

Another important challenge is the DNA sample requirements for library preparation and sequencing. Since the PacBio sequences single molecules of DNA directly and does not typically employ an amplification step, library construction requires 1 to 10 μg of high-quality purified DNA as starting material, depending on the intended application and library insert size. (A 10- to 20-kb insert library typically requires close to 10 μg without contaminants or damage.) In many situations, such large amounts of purified high-quality DNA may not be feasible, and because of the read lengths involved, whole-genome amplification techniques are often not an option.

Output limitations

Relative to short-read sequencers, the PacBio RSII has limited value for certain applications, such as the sequencing and analysis of complex metagenomics samples, since the runs are relatively costly and the number of reads produced is far lower. That said, the long read lengths produced by the PacBio are an important advantage in the phasing and assembly of metagenomics samples and may add critical value for less complex metagenomic samples, or with the improved output from the newer Sequel instrumentation.

Multiplexing

Multiplexing, or the ability to run multiple samples on a single sequencing run, remains challenging on the PacBio RSII and has yet to be fully developed and validated for the Sequel. Currently, most laboratories are running single isolates per SMRTcell, which results in exceptional draft genome assemblies but rapidly becomes cost prohibitive at scale. The lack of established multiplexing methods is an important limitation to microbial sequencing on the newer Sequel instruments, which have significantly higher output that cannot be applied to multiple samples. Pacific Biosciences recently released official protocols and consumables for multiplexing samples on the Sequel instruments, with widespread availability during Q2 2018.

Nanopore sequencing (Oxford Nanopore)

Another relatively new long-read sequencing technology is often referred to as nanopore sequencing, since it relies on engineered protein nanopores to linearize and sequence molecules of nucleic acid. The Oxford Nanopore MinION was introduced in 2012, with a preview program that launched for early adopters in 2014. Much like for PacBio RS, the first iterations of the MinION platform were beset by high rates of error. Recent improvements in the hardware, chemistry, nanopore configuration, and base calling algorithms have all greatly

improved the accuracy and throughput of the MinION, with the release of the MinION Mk 1B in 2016.

Oxford Nanopore's implementation of nanopore sequencing relies on an array of engineered protein nanopores embedded in a synthetic polymer membrane. During library construction, the template DNA is either sheared to a specific size range, PCR amplified, or left intact, depending upon the specific sequencing application. In most protocols, the double-stranded DNA strands are repaired and a paired set of sequencing adapters and hairpin adapters are ligated onto each end of the template, such that when the strand is denatured and linearized for sequencing, the template sense strand will be sequenced, followed by the hairpin adapter sequence and, finally, the antisense template strand. This is known as a two-direction read. The linearized single strand is fed through the protein nanopore by a motor protein, which binds to the end of the double-stranded template, unzips the molecule, and ratchets it through the central channel of the pore, which is several nanometers in diameter. A voltage potential is applied across the membrane. As the linearized, single-stranded DNA molecule is fed through the nanopore channel by the motor protein, the bases disrupt the current within the channel. By measuring these local changes in resistivity across each individual nanopore, the nanopore is able to base call the sequence in 2-mer or 3-mer increments.

Nanopore sequencing has shown important promise for bacterial genome assembly and real-time metagenomics, and the MinION is a particularly compelling platform for a number of reasons. The first is its diminutive size, which makes NGS in remote field locations a practical option for infectious disease surveillance, diagnostics, and public health research. For example, the MinION has recently been used for sequencing of both Ebola and Zika viruses, with reference and surveillance sequencing applications both in the laboratory and in the field (20, 21). The second is cost: at a fraction of most short- and long-read sequencing platforms, nanopore sequencing is an increasingly feasible option for smaller laboratories and those that cannot justify or support sustained investment in large capital instrumentation. Nanopore sequencing is also significantly faster than many other sequencing approaches since instrument cycling is not required and signal data can be streamed directly from the instrument. This streaming model enables rapid base calling and real-time analysis, allowing users to "read until" sufficient data have been gathered and to continuously assess the progress and quality of data generation, rather than analyzing the entire data set after the completion of a run. Engineered protein nanopores also offer remarkable flexibility in terms of format, capability, and scalability and have been demonstrated to support direct sequencing of other types of complex biomolecules in addition to DNA, including RNA and peptides (22).

In addition to existing systems from Oxford Nanopore, a number of other nanopore-based platforms are currently under development or in the premarket space, including an integrated circuit-based platform from Roche Genia (https://sequencing.roche.com/en/technology-research/technology/nanopore-sequencing.html) and several early-phase proof-of-concept platforms from academic laboratories (23). As nanopore-based sequencing technologies continue to mature, and other systems enter the market, these flexible, cost-effective sequencing instruments will almost certainly play an increasing role in microbiological testing, particularly in small laboratories, in clinics, and in the field.

MICROBIAL NEXT-GENERATION SEQUENCING BIOINFORMATIC WORKFLOWS

Next-generation sequencing technologies, particularly short-read technologies, can be used across a wide range of different applications in order to derive actionable public health information. For most public health applications having to do with infectious disease surveillance and outbreak response, comparative microbial genomics is an important first step in understanding the molecular epidemiology, risk factors, and transmission dynamics of the pathogen. In general, the analysis of microbial genomes may be broadly categorized into (i) *de novo* sequencing approaches, (ii) those based on reference mapping, and, finally, (iii) assembly-free methods. The following section describes each of these methods, briefly, in turn.

Since the advent of NGS and other high-throughput laboratory technologies, the number and sophistication of open-source bioinformatics tools, scripting languages, and workflow management frameworks have grown exponentially. While this openness has fostered rapid innovation and improvement, many bioinformatic tasks have been approached using multiple algorithmic approaches or iterative implementations. It is therefore incumbent upon bioinformaticians, laboratorians, and epidemiologists alike to fully understand the strengths, limitations, biases and assumptions of their bioinformatic tools and pipelines and to take these elements into consideration when developing and validating new analytical workflows, and in interpreting the resulting

output in the context of an investigation. For example, single nucleotide polymorphisms (SNPs) are often used as a means to assess the phylogenetic relationships between bacterial pathogens. An NGS SNP typing pipeline consists of multiple procedural steps, including (i) quality assessment and trimming, (ii) reference selection and masking, (iii) read mapping, (iv) variant calling, (v) filtering, (vi) phylogenetic analysis, (vii) tree building, and (viii) visualization and reporting. Each one of these steps may be accomplished somewhat interchangeably by one or more different bioinformatic tools or algorithmic approaches, each of which carries different parameterizations, input requirements, and assumptions. As a result, complex bioinformatic pipelines can vary quite significantly in their composition, although in many cases they may give highly concordant results. It is therefore important for the end user to understand what the pipeline is doing and why before making any meaningful interpretations of the output.

Pipelines for Microbial Comparative Genomics

Regardless of the downstream bioinformatic approach, most microbial analysis pipelines include some common steps to identify and correct for issues with sequence quality. This includes an initial assessment of the overall quality and performance metrics for each sequencing run, demultiplexing of the sample into individual isolate readsets, if applicable, read level quality control, and trimming/filtering poor quality and adapter sequences.

De novo assembly

De novo assembly attempts to assemble each microbial genome without any *a priori* knowledge of the underlying genomic structure: reads are assembled based on the overlap and scaffolding data at hand. Because *de novo* assembly does not rely on a defined reference genome, it is often the preferred approach for poorly characterized organisms or those with high levels of genomic plasticity and extrachromosomal sequence, since it may assemble all of these into useful contigs. Therefore, *de novo* assembly may prove useful in the analysis of highly divergent organisms or those with significant structural differences or for the simultaneous analysis of chromosomal and extrachromosomal (e.g., plasmid) sequences. However, this lack of reference context may also be an important disadvantage, particularly in the resolution of complex genomic structure and, in particular, compound repeat regions, which often do not assemble cleanly from short-read sequence data alone. Long-read sequence data, such as those from Pacific Biosciences or Nanopore instruments, may be used to overcome these assembly limitations and often result in well-supported single-contig assemblies of target microbial genomes. While a number of different denovo assemblers have been developed, SPAdes (http://cab.spbu.ru/software/spades/) is among the most common for microbial *de novo* sequence assembly. In addition to assembly, *de novo* pipelines may also include other steps, such as whole-genome alignment for genomic comparison, using tools such as *mauve* and *harvest*; annotation and functional prediction, using tools such as *prokka*; tree building, using a variety of tools and algorithms; and a range of comparative genomic analysis, including MLST and genotypic prediction of antimicrobial resistance, using tools such as *arbricate*.

Reference mapping

In reference mapping or guided assembly, a closely related reference genome is used as a template for mapping reads from the query isolate. Reads that are mapped against the reference may indicate important functional differences between the query and reference sequences, and this variant analysis is the basis for SNP typing and other genomic strain typing techniques. Reference mapping works well for organisms with relatively low population diversity or with a carefully selected reference isolate, where the mapping of reads will not be impacted by major structural differences, evolutionary divergence, or the introduction or loss of extrachromosomal sequences. These approaches are relatively computationally fast but may not be appropriate for highly diverse sequences or for the analysis of mobile genetic elements and linear reference sequences. Pipelines that implement reference mapping often include alignment (bwa and bowtie2) and variant calling (gatk, varscan, and freebayes) as core functions, with variant filtering, annotation (breseq and prokka), comparative genomics, phylogenetics, and genotypic prediction of antimicrobial resistance and virulence characteristics.

Assembly-free methods

Assembly-free methods are relatively recent algorithmic approaches for applied comparative genomics and are often used in conjunction with both reference mapping and *de novo* assembly. As their name implies, assembly-free methods typically use k-mer or hash-based approaches to analyze query sequences based on the composition of raw sequence reads and are not dependent on computationally demanding sequence assembly or mapping steps. For this reason, they tend to be extremely fast to compute, often at the expense of accuracy, and make an ideal first-pass analysis or quality

assessment of query sequences. These k-mer-based approaches may also be used for rapid molecular epidemiologic analyses, including MLST (e.g., srst2) and SNP-based methods (e.g., kSNP) (24, 25), as well as the genotypic prediction of phenotypic traits (e.g., k-mer resistance) (26).

Pipeline and Workflow Management

A number of different open-source and commercial software packages have been developed to facilitate microbial bioinformatics and workflow management. Popular bioinformatic pipelines for microbial sequence analysis have been developed using Galaxy (https://www.usegalaxy.org), CLCbio Genome Workbench, and BioNumerics, among others. These platforms provide a graphical environment for sample, workflow, and parameter management and provide a consistent framework for protocol validation, standardization, and quality assessment that does not necessarily require a high degree of technical bioinformatic capacity from the end user. For this reason, large-scale molecular surveillance systems, such as PulseNet (http://www.cdc.gov/pulsenet), and newer initiatives, such as IRIDA (http://www.irida.ca), rely heavily on these workflow and sample management systems to deploy complicated bioinformatic analysis pipelines for routine public health use (27).

CHALLENGES OF IMPLEMENTING NGS IN CLINICAL AND PUBLIC HEALTH MICROBIOLOGY

Until the commercial introduction of NGS platforms in the mid-2000s, the cost and complexity of genetic sequencing limited its use as a routine, front-line technique in most clinical and public health laboratories. The tremendous impact of accessible, high-throughput sequencing since then cannot be easily overstated, and it represents a fundamental change in laboratory capabilities, standard practice, workflow management, and staffing requirements. As the cost and technical barriers to NGS implementation continue to decrease, sequence-driven methods are rapidly replacing or enhancing traditional microbiologic approaches, requiring significant adaptation and change. The advantages of NGS are often offset by important technical and logistical challenges, including rapidly increasing demand for laboratory and bioinformatics capacity, the lack of standard operating protocols, the need for rapid workforce realignment and training, new requirements for data management, analysis, and sharing, complexities involved with clinical reporting, and gaps or inconsistencies in existing bioinformatics reference databases and other critical resources.

Data Volume and Complexity

The transition to genomics-based methods represents a significant increase in the volume and complexity of the data that must be managed, analyzed, and interpreted. Compared to conventional molecular methods for strain typing and characterization (e.g., PCR and pulsed-field gel electrophoresis or multilocus variable number tandem repeat analysis), the raw data from NGS may represent an increase in volume of 10- to 10,000-fold. A small outbreak investigation involving several dozen bacterial isolates might easily surpass 25 gigabytes of raw sequence data, which is roughly the same volume of data that the entire PulseNet USA database would accumulate over an entire year before the advent of NGS. As the PulseNet system transitions to whole-genome sequencing and whole-genome MLST-based typing, the raw data generated by the CDC and its partners could easily surpass 100 terabytes per year. At this rate of sequence generation, data transmission, management, and storage all become important considerations for large-scale molecular surveillance, as do the bioinformatic approach and policies for data sharing and retention (27, 28).

Overall, the scale and complexity of pathogen genomic sequence data also represent an important challenge for data management and integration, particularly in linking important microbial findings with their respective epidemiologic context. In many cases, bioinformatic analysis of microbial sequence data is no longer the primary bottleneck to public health surveillance and outbreak response: integrating laboratory and epidemiologic data often presents a significantly greater challenge. As bioinformatic pipelines and laboratory data systems continue to evolve, the development of strategies to integrate laboratory and epidemiologic data quickly and accurately and to share these findings across multidisciplinary teams will be increasingly essential to an effective and timely public health response.

Pace of Technological Change

The technology space around NGS has evolved rapidly, with frequent iterations of the instrumentation, sequencing reagents, and consumables, punctuated every few years by new platform introductions. In individual laboratories, this rate of change presents important challenges for capital equipment planning and workforce development, and for public health, it represents an

important obstacle to the development and validation of standardized assays and tools across networks of laboratories. With new instrumentation, and constant change in the reagents and consumables across the entire NGS workflow, capital equipment costs, and the need for ongoing revision, validation and qualification of routine assays to meet the current standards of practice may prove to be a difficult and ongoing challenge.

To compound matters, many public health departments overlook or underestimate the need for investments in information technology (IT), including high-speed networking, storage, computation, and high-speed connectivity. Often, these costs may significantly outweigh the costs of the laboratory infrastructure itself. Moreover, in many public health departments, IT staff are not adequately resourced to support these new scientific computing requirements, and both outsourcing and access to cloud computing resources may be limited by cost, security, or legal considerations.

Lack of High-Performance Computing Resources and Bioinformatics Expertise

Genomics and other high-throughput laboratory technologies have greatly increased the demand for bioinformatics and data science expertise in public health. Recruiting skilled bioinformaticians to clinical and public health careers presents a significant challenge, however, since these skills are also in high demand in academia and the private sector. It is not always possible for public health departments to compete for these individuals on the basis of salary alone, even with a compelling mission, interesting technical challenges, and a wealth of opportunities. A more practical strategy may be to provide the necessary training and resources to bench microbiologists, epidemiologists, and clinical laboratory staff to enable them to become competent users of bioinformatic tools and to realign existing training and fellowship programs to prioritize the necessary workforce skills.

Infrastructure is another concern. Most clinical and public health laboratories do not have local access to the necessary scientific and high-performance computing resources that are needed for complex bioinformatic analyses. While cloud computing may be feasible in some jurisdictions, for many laboratories, information security and patient privacy considerations can pose additional challenges. The technical and computational complexity of analytical tasks can vary widely, and therefore, public health laboratories need both (i) simple, push-button analytics for routine and standardized applications and (ii) access to flexible, cost-effective advanced bioinformatics for more specialized applications. Delivering accessible and sustainable bioinformatic skills and infrastructure for both of these requirements will continue to be an important challenge.

Data and Metadata Sharing

The responsible sharing of pathogen genomic data and metadata marks another important challenge, particularly in the context of public health. In the past, most surveillance and outbreak response data were collected and analyzed in secure databases. Today, a growing number of laboratory-based surveillance systems release microbial sequence data into the public domain in near-real time—a significant departure from traditional data sharing models. This data openness underscores a growing public health commitment to open data and is intended to accelerate diagnostic development, to support basic research and collaboration, and to enable ongoing collaboration in the assessment of critical genotypic markers and changes in microbial population dynamics.

Any release of genomic sequence data from patient isolates must balance patient privacy considerations against the utility and value of public data release. While high-throughput sequencing platforms have begun to converge on some common data standards and interchange formats (e.g., FASTQ) for sequence data and corresponding quality scores, standards for minimum associated metadata are still under active development and review. Among the efforts currently underway, a number of national public health authorities are working with public data resources, such as the National Center for Biotechnology Information (NCBI), to develop common and machine-readable metadata standards for sample descriptions, basic demographics and outcomes, and functional characteristics such as antimicrobial resistance and virulence (28).

Lack of Definitive Reference Data and Standardized Methods

Databases of high-quality, curated microbial sequences are increasingly essential for comparative pathogen genomics, sequence-based strain typing, and molecular characterization. While public sequence repositories, such as NCBI, continue to grow exponentially and provide an important basis for curated reference databases, sequences must be carefully reviewed for quality, completeness, representativeness, and correctness. Expert curation is vital, and many recalcitrant species, commensal organisms, and rare or unconventional pathogens remain critically underrepresented in even the most

comprehensive sequence databases. In practical terms, these gaps or uneven representation in reference databases can contribute to critical errors in downstream analyses, including misidentification and mischaracterization of sample queries. A number of large-scale sequencing efforts are currently under way to help improve the quality and comprehensiveness of microbial reference sequences. Expert-curated databases such as CDC's MicrobeNet (https://microbenet.cdc.gov) are useful, particularly for rare, unusual, or underrepresented species, and presenting genomic sequences in the context of their corresponding genotypic, phenotypic, and proteomic features enables complex queries and identification based on the data at hand.

Limits to Phenotypic Prediction and the Impact of Culture-Independent Diagnostic Testing

Although NGS-based microbial identification and characterization methods can provide a relatively complete and comprehensive profile of both known and unknown microbial pathogens, orthogonal or confirmatory testing is often critical to definitively assess features such as virulence and antimicrobial susceptibility. Predicting the antimicrobial resistance phenotype of an organism directly from genomic data is increasingly feasible, particularly for acquired resistance genes (28, 29). However, in some instances, microbial genotype and phenotype may not correlate perfectly due to factors that are not reflected in the genome of the organism (e.g., point mutations, altered gene expression, epigenetic factors, or differences in genomic structure), particularly for virulence and resistance mechanisms that are multifactorial or incompletely understood.

The limitations of genotypic methods are an important consideration, particularly as clinics and hospital laboratories transition to culture-independent diagnostic testing and point-of-care molecular testing based on syndromic presentation. Traditional public health surveillance and outbreak response relies heavily on the availability of culture isolates for more detailed testing, including molecular strain typing and characterization. As the availability of clinical cultures begins to decrease, the effectiveness of public health surveillance and outbreak response will be significantly impaired, without corresponding advances in public health laboratory technology. One strategy that is currently being explored is the direct or targeted metagenomic analysis of patient specimens. This sequence-based method of characterizing pathogens directly from patient samples is complicated and presents significant technical challenges. Moreover, it may not provide a complete set of actionable data, if functional susceptibility and other phenotypic characteristics cannot be reliably predicted (28, 30).

As the impact of NGS and other high-throughput laboratory technologies continues to expand, many clinical and public health microbiology laboratories are undergoing fundamental changes in their workflow and capabilities. The increased speed, accuracy, and resolution of NGS over conventional molecular diagnostic methods are already making significant contributions to surveillance, outbreak detection, and public health response. The development of consensus standards, high-quality reference databases, laboratory workforce capacity, and standardized analytical approaches is critical to sustained implementation of NGS and other advanced laboratory technologies. While many challenges remain, these technologies represent an important new era in molecular epidemiology and herald unprecedented developments in molecular surveillance and the study of infectious diseases.

DISCLAIMER

Specific vendors, system names, and instrument models are included for informational purposes only. Their use does not imply endorsement by the author, the Department of Health and Human Services, or the Centers for Disease Control and Prevention.

CURATED COLLECTION

Click here to read other articles in this collection.

REFERENCES

1. Oster AM, Wertheim JO, Hernandez AL, Ocfemia MCB, Saduvala N, Hall HI. 2015. Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. *J Acquir Immune Defic Syndr* **70:**444–451. http://dx.doi.org/10.1097/QAI.0000000000000809.

2. Cookson BD, Robinson DA, Monk AB, Murchan S, Deplano A, de Ryck R, Struelens MJ, Scheel C, Fussing V, Salmenlinna S, Vuopio-Varkila J, Cuny C, Witte W, Tassios PT, Legakis NJ, van Leeuwen W, van Belkum A, Vindel A, Garaizar J, Haeggman S, Olsson-Liljequist B, Ransjo U, Muller-Premru M, Hryniewicz W, Rossney A, O'Connell B, Short BD, Thomas J, O'Hanlon S, Enright MC. 2007. Evaluation of molecular typing methods in characterizing a European collection of epidemic methicillin-resistant *Staphylococcus aureus* strains: the HARMONY collection. *J Clin Microbiol* **45:**1830–1837. http://dx.doi.org/10.1128/JCM.02402-06.

3. Chen F-J, Hiramatsu K, Huang I-W, Wang C-H, Lauderdale T-LY. 2009. Panton-Valentine leukocidin (PVL)-positive methicillin-susceptible and resistant *Staphylococcus aureus* in Taiwan: identification of oxacillin-susceptible *mecA*-positive methicillin-resistant *S. aureus*. *Diagn Microbiol Infect Dis* **65:**351–357. http://dx.doi.org/10.1016/j.diagmicrobio.2009.07.024.

4. Denisuik AJ, Lagacé-Wiens PR, Pitout JD, Mulvey MR, Simner PJ, Tailor F, Karlowsky JA, Hoban DJ, Adam HJ, Zhanel GG, Canadian Antimicrobial Resistance Alliance. 2013. Molecular epidemiology of extended-

spectrum β-lactamase-, AmpC β-lactamase- and carbapenemase-producing *Escherichia coli* and *Klebsiella pneumoniae* isolated from Canadian hospitals over a 5 year period: CANWARD 2007-11. *J Antimicrob Chemother* **68**(Suppl 1)**:**i57–i65. http://dx.doi.org/10.1093/jac/dkt027.

5. Chen L, Cai Y, Zhou G, Shi X, Su J, Chen G, Lin K. 2014. Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens. *PLoS One* **9:**e88886. http://dx.doi.org/10.1371/journal.pone.0088886.

6. Centers for Disease Control and Prevention. *Streptococcus pyogenes emm* sequence database. https://www2a.cdc.gov/ncidod/biotech/strepblast.asp.

7. Nadon CA, Trees E, Ng LK, Møller Nielsen E, Reimer A, Maxwell N, Kubota KA, Gerner-Smidt P, Collective the MLVA Harmonization Working Group. 2013. Development and application of MLVA methods as a tool for inter-laboratory surveillance. *Euro Surveill* **18:**20565. http://dx.doi.org/10.2807/1560-7917.ES2013.18.35.20565.

8. Fawley WN, Knetsch CW, MacCannell DR, Harmanus C, Du T, Mulvey MR, Paulick A, Anderson L, Kuijper EJ, Wilcox MH. 2015. Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for *Clostridium difficile*. *PLoS One* **10:**e0118150. http://dx.doi.org/10.1371/journal.pone.0118150.

9. MacCannell D. 2016. Next generation sequencing in clinical and public health microbiology. *Clin Microbiol Newsl* **38:**169–176. http://dx.doi.org/10.1016/j.clinmicnews.2016.10.001.

10. Gargis AS, Kalman L, Bick DP, da Silva C, Dimmock DP, Funke BH, Gowrisankar S, Hegde MR, Kulkarni S, Mason CE, Nagarajan R, Voelkerding KV, Worthey EA, Aziz N, Barnes J, Bennett SF, Bisht H, Church DM, Dimitrova Z, Gargis SR, Hafez N, Hambuch T, Hyland FC, Luna RA, MacCannell D, Mann T, McCluskey MR, McDaniel TK, Ganova-Raeva LM, Rehm HL, Reid J, Campo DS, Resnick RB, Ridge PG, Salit ML, Skums P, Wong LJ, Zehnbauer BA, Zook JM, Lubin IM. 2015. Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* **33:**689–693. http://dx.doi.org/10.1038/nbt.3237.

11. Loman NJ, Pallen MJ. 2015. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol* **13:**787–794. http://dx.doi.org/10.1038/nrmicro3565.

12. Illumina, Inc. 2016. Illumina two-channel SBS sequencing technology. https://www.illumina.com.

13. Davies K. 27 September 2011. Powering preventative medicine. *Bio-IT World*. http://www.bio-itworld.com/issues/2011/sept-oct/powering-preventative-medicine.html.

14. Life Technologies Corporation. 2011. Application note: Ion Torrent amplicon sequencing. https://www.appliedbiosystems.com.

15. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. 2012. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10:**599–606. http://dx.doi.org/10.1038/nrmicro2850.

16. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, 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. http://dx.doi.org/10.1126/science.1150427.

17. Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* **27:**847–850. http://dx.doi.org/10.1038/nbt.1561.

18. Hoffmann M, Muruvanda T, Allard MW, Korlach J, Roberts RJ, Timme R, Payne J, McDermott PF, Evans P, Meng J, Brown EW, Zhao S. 2013. Complete genome sequence of a multidrug-resistant *Salmonella enterica* serovar typhimurium var. 5– strain isolated from chicken breast. *Genome Announc* **1:**e01068-13. http://dx.doi.org/10.1128/genomeA.01068-13.

19. Nandi T, Holden MT, Didelot X, Mehershahi K, Boddey JA, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. 2015. *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res* **25:**129–141. http://dx.doi.org/10.1101/gr.177543.114.

20. Kilianski A, Roth PA, Liem AT, Hill JM, Willis KL, Rossmaier RD, Marinich AV, Maughan MN, Karavis MA, Kuhn JH, Honko AN, Rosenzweig CN. 2016. Use of unamplified RNA/cDNA-hybrid nanopore sequencing for rapid detection and characterization of RNA viruses. *Emerg Infect Dis* **22:**1448–1451. http://dx.doi.org/10.3201/eid2208.160270.

21. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, Nair S, Neal K, Nye K, Peters T, De Pinna E, Robinson E, Struthers K, Webber M, Catto A, Dallman TJ, Hawkey P, Loman NJ. 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol* **16:**114. http://dx.doi.org/10.1186/s13059-015-0677-2.

22. Feng Y, Zhang Y, Ying C, Wang D, Du C. 2015. Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinformatics* **13:**4–16. http://dx.doi.org/10.1016/j.gpb.2015.01.009.

23. Stranges PB, Palla M, Kalachikov S, Nivala J, Dorwart M, Trans A, Kumar S, Porel M, Chien M, Tao C, Morozova I, Li Z, Shi S, Aberra A, Arnold C, Yang A, Aguirre A, Harada ET, Korenblum D, Pollard J, Bhat A, Gremyachinskiy D, Bibillo A, Chen R, Davis R, Russo JJ, Fuller CW, Roever S, Ju J, Church GM. 2016. Design and characterization of a nanopore-coupled polymerase for single-molecule DNA sequencing by synthesis on an electrode array. *Proc Natl Acad Sci U S A* **113:**E6749–E6756. http://dx.doi.org/10.1073/pnas.1608271113.

24. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, Zobel J, Holt KE. 2014. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6:**90. http://dx.doi.org/10.1186/s13073-014-0090-6.

25. Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics* **31:**2877–2878. http://dx.doi.org/10.1093/bioinformatics/btv271.

26. Clausen PTLC, Zankari E, Aarestrup FM, Lund O. 2016. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J Antimicrob Chemother* **71:**2484–2488. http://dx.doi.org/10.1093/jac/dkw184.

27. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, CDC PulseNet Task Force. 2001. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis* **7:**382–389. http://dx.doi.org/10.3201/eid0703.017303.

28. Gwinn M, MacCannell DR, Khabbaz RF. 2016. Integrating advanced molecular technologies into public health. *J Clin Microbiol* **55:**703–714.

29. McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, Folster JP, Ayers SL, Lam C, Tate HP, Zhao S. 2016.Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrob Agents Chemother* **60:**5515–5520. http://dx.doi.org/10.1128/AAC.01030-16.

30. Cronquist AB, Mody RK, Atkinson R, Besser J, Tobin D'Angelo M, Hurd S, Robinson T, Nicholson C, Mahon BE. 2012. Impacts of culture-independent diagnostic practices on public health surveillance for bacterial enteric pathogens. *Clin Infect Dis* **54**(Suppl 5)**:**S432–S439. http://dx.doi.org/10.1093/cid/cis267.