Next Generation Sequencing for the Detection of Foodborne Microbial Pathogens

14

Travis G. Wentz, Lijun Hu, Thomas S. Hammack, Eric W. Brown, Shashi K. Sharma, and Marc W. Allard

14.1 Introduction

The rapid detection and typing of DNA belonging to known and emerging pathogens represents one of the most fundamental and frequently encountered tasks by state and national health laboratories. Over the past decade, next generation sequencing (NGS) platforms have been incorporated into a range of public health programs responsible for surveilling, detecting, and investigating/responding to infectious disease outbreaks. NGS has been rapidly integrated into the field of pathogenic foodborne microbiology as both a primary and supportive detection tool and is routinely used in the analysis of isolates from many prominent foodborne bacterial pathogens including *Salmonella*, *Listeria*, *Escherichia coli*, *Shigella*, and neurotoxigenic *Clostridium*. Collectively, 31 major foodborne pathogens are estimated to result in 9.4 million instances of illness leading to 55,961 hospitalizations and 1351 deaths per year in the United States; figures dwarfed by estimates over the same period for unspecified agents responsible for 38.4 million cases of acute gastroenteritis, 473,832 hospitalizations, and 5072 deaths [1, 2]. The relatively well-defined and studied major foodborne pathogens often are associated with established regulatory procedures for their detection and verification and are often the focus of major public health programs. A number of these bacteria have been the subject of large multi-center NGS-enabled whole genome sequencing (WGS) initiatives, which have begun to fundamentally change the landscape of disease surveillance. While a diversity of factors is responsible for the many cases of acute gastroenteritis with unspecified etiology, the possibility exists that a substantial

T. G. Wentz · L. Hu · T. S. Hammack · E. W. Brown · S. K. Sharma · M. W. Allard (✉)
Division of Microbiology, Office of Regulatory Science, Center for Food Safety and Applied Nutrition, Food and Drug Administration, College Park, MD, USA
e-mail: Travis.Wentz@fda.hhs.gov; Lijun.Hu@fda.hhs.gov; Thomas.Hammack@fda.hhs.gov; Eric.Brown@fda.hhs.gov; Shashi.Sharma@fda.hhs.gov; Marc.Allard@fda.hhs.gov

S. K. Singh, J. H. Kuhn (eds.), *Defense Against Biological Attacks*,
https://doi.org/10.1007/978-3-030-03071-1_14

number are attributable to uncharacterized, cryptic, or conditional pathogens that currently evade identification. The steady growth in WGS and metagenomic sequence data from pathogenic and non-pathogenic organisms already has provided critical insight into horizontally mobile genomic elements and revealed that some critical virulence factors may have a broader distribution than was previously understood.

NGS is a transformative technology, and the sequence data produced by NGS are impacting the field of pathogen detection in profound ways. This chapter explores what NGS platforms are, the types of sequence data they can produce, and how sequence data are being leveraged to enhance the detection of foodborne bacterial pathogens. In the first part of the chapter we begin with a brief history of the emergence of NGS technology and its early integration into the field of bacterial pathogenesis. Next, we provide an overview of core concepts used in the preparation of nucleotide data for whole genome sequencing before transitioning into an overview of several commonly encountered NGS platforms and the state of the sequence data that is produced by them.

In the second half, we focus on the utilization of NGS data as a tool for pathogen typing and detection. The process by which viruses, organisms, and/or their toxic factors drive pathogenesis can be immensely complex and diverse. Although WGS can be highly complementary to pathogen detection goals, there is no one-size-fits-all answer for how to utilize WGS data. Options will differ based on a range of factors that are often specific to the organism genome being sequenced. To provide a broad overview we discuss gnome assembly and the applications of NGS data in the context of two foodborne pathogens, *Salmonella enterica* (*S. enterica*) and *Clostridium botulinum* (*C. botulinum*). These two organisms differ substantially in regard to underlying biology, disease outbreak frequency, pathogenesis, and detection goals. These differences allow us to explore the immense variety of options an investigator faces once WGS data are acquired, highlight how these data can be applied to detection goals, and demonstrate how these goals can vary from organism to organism. We explore (1) the use of WGS as a high-resolution molecular typing tool, (2) its compatibility with other typing schemes, (3) and ways to utilize the data encoded within the genome to detect and explore virulence factors.

14.2 Next Generation Sequencing: Background and History

The most approachable use of NGS technology lies in the ability of most modern platforms to rapidly and accurately produce WGS data. For culturable bacteria, modern NGS platforms can generate WGS data suitable for de novo genome assembly or resequencing purposes within a timeframe of several hours to several days [3]. Reference-based resequencing approaches map sequenced reads to an already existing genome assembly, usually for purposes of variant detection or rapid typing as part of an established bioinformatics workflow. De novo sequencing is computationally driven, direct assembly of sequenced reads into larger contiguous

sequences. Although conceptually distinct, these applications are not mutually exclusive, and resequencing is often used following de novo assembly as part of an iterative approach to acquire a more accurate, consensus assembly. Most NGS platforms perform WGS by utilizing a random sequencing approach to generate large quantities of sequenced reads that are later computationally reassembled into a contiguous sequence, or contig. The value of this approach was demonstrated when investigators at The Institute for Genomic Research utilized random sequencing to assemble the complete bacterial genome of the pathogen *Haemophilus influenzae* in 1995 [4]. WGS data produced by most NGS platforms require a similar process of post-run assembly of sequenced reads. Pre-NGS, WGS projects relied predominantly on automated capillary sequencing or automated Sanger/ddNTP sequencing.

Though the 1990s and early 2000s, time and labor costs associated with preparing clone libraries for sequencing kept genome scale sequencing projects out of reach for most laboratory groups that were not solely dedicated to genomic projects. Most modern NGS platforms are optimized in ways that substantially simplify the library preparation steps prior to sequencing. The massively parallel nature and increased processing ability of NGS technologies enables accurate, high quality sequencing to occur in a timeframe of days or weeks, instead of the months or years required for earlier projects. Compared to automated Sanger/ddNTP sequencing, these platforms are fast, cost effective, and accessible to researchers in a range of settings. There is some sequence accuracy tradeoff, which we discuss as part of our overview of several sequencing platforms, and there can be challenges present in the storage, analysis, and effective utilization of produced data. However, the approachability of NGS has resulted in its rapid integration as an indispensable tool in the field of pathogen detection.

The first standalone next generation sequencer, the Life Sciences 454 pyrosequencer, became available in 2004. By 2008 the field consisted of a diverse collection of platforms, many operating in fundamentally different ways but all achieving the common task of producing DNA sequence data through means faster and cheaper than previous options [5, 6]. As an emerging technology, NGS adoption was not immediate. Well into the start of the NGS era, the whole genome of the *Bacillus anthracis* Ames Ancestral strain, utilized as part of the comparative genome study linked to the 'Amerithrax' investigation, was sequenced via automated capillary sequencing [7, 8]. Nonetheless, the 'Amerithax' case demonstrated the power of whole genomic data to resolve sequence differences between very closely related strains and the potential application of bacterial WGS data as an investigatory tool. Even low coverage WGS via capillary sequencing is cost prohibitive for most sequencing projects. As NGS products continue to mature as a technology, specialized pathogen oriented datasets that combine detailed isolate collection metadata, isolate WGS, and rapid phylogenetic analysis have been developed and are beginning to shape the future of pathogen detection, outbreak response, and source tracking.

While most modern sequencing platforms provide fast and accurate data, there are several important considerations when selecting a platform depending on

intended purpose. Key factors for WGS include genome size, sequencing coverage, the number of cultured isolates that require sequencing, the nature of data generated by the sequencing platform, and the availability of computing resources for assembling and interpreting the whole genomic data. Table 14.1 provides a brief overview of some of the most frequently encountered NGS platforms in use today. Each NGS platform/family of platforms determines sequence through substantially different means and prior to sequencing, extracted DNA is subjected to a library preparation step which ensures the DNA is physically structured in a way that is conducive to platform operation. Broadly, short-read sequencers such as the Illumina MiSeq, NextSeq, and MiniSeq platforms generate copious quantities of short reads that can be mapped against reference sequences or used to create de novo assemblies that assemble into multiple contigs. Long-read sequencers, including the Pacific Biosciences (PacBio; Menlo Park, California, US) and Oxford Nanopore platforms (Oxford, UK), can produce long reads for scaffolding genomic regions together and with sufficient coverage depth, produce de novo assemblies consisting of a closed genome. We discuss in detail the types of output that can be expected from the platforms covered in the table but first provide a brief introduction to some of the core concepts necessary for conducting a successful sequencing run.

14.3 Core Concepts in Sequencing

The laboratory wet work process of sequencing the genome of an organism for purposes of bacterial WGS often follows a set path. A bacterial isolate of interest is cultured on solid media and grown to desired cell density. Genomic DNA is extracted, quantified, prepared as a genomic library, and ultimately input into a sequencing platform which generates sequenced nucleotide reads as an output. Although we will describe these concepts assuming a typical WGS protocol, many are general terms that are broadly applicable to many types of sequence output.

14.3.1 Genome Size and %GC Content

Genome size, the sum of DNA in an organism measured in nucleotide basepairs (bp) varies substantially across and within the major domains of life. As of August 2017, complete genomes in NCBI/Genbank ranged in length for viruses, including viroids, from 220 to 2,473,870 bases and for prokaryotes from 112,031 to 16,000,000 bp. In the case of viral and prokaryotic genomes, genome sizes and the amounts of coding DNA increase in a mostly proportional linear fashion. On the other hand, eukaryotic genomes are extremely variable in size and often contain large amounts of non-transcribed DNA with regulatory or unknown functions [9]. For smaller eukaryotic genomes, e.g. those of some yeasts, WGS with the intention of de novo assembly may be a readily accessible but there are additional complexities by the potentially diploid+ nature of these organisms [10]. Special

Table 14.1 Sequencing statistics associated with assemblies in Fig. 14.1

Platform	# Reads	Mean Read Length (bp)	Mean coverage	Contigs	N50 (bp)	Estimated genome length (bp)	Notes
Illumina MiSeq	903,570	250 (paired, mean distance 325 bp)	40×	174	45,828	4,334,787	Preparation via Nextera XT Kit; Sequenced in a normalized run alongside 23 other bacterial samples
PacBio RS-II	40,563 reads, 142,992 subreads	11,022	250×	1	N/A	4,501,946	10-kbp size selected library

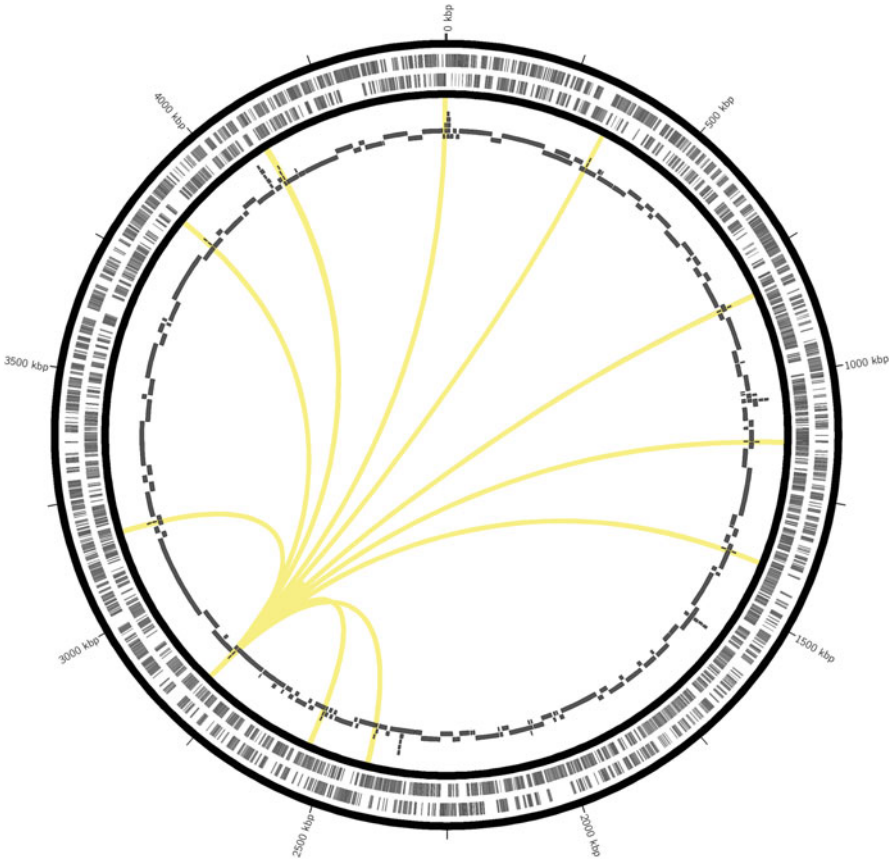

Fig. 14.1 Circos plot of a 4.5-Mbp bacterial chromosome. Solid black bands represent a closed annotated genome assembly generated from long-read sequencing on a Pacbio platform. Fragmented inner bands are sequences from a contig MiSeq short-read assembly, locally aligned against the closed genome. Yellow bands highlight the location of identical IS21 insertion sequences throughout the genome, discussed in Fig. 14.2

consideration of sequencing platform, data capacity, and bioinformatic support should be taken for projects aimed at sequencing eukaryotic pathogens such as protozoa and fungi with genomes longer than 20–25 Mbp. Figure 14.1 provides a model visualization of a closed bacterial genome with the length of 4,500,000 bp that is fairly typical of an average bacterial genome. An additional feature, %GC nucleotide content should also be considered prior to sample preparation as some sequencing platforms can have markedly different error profiles depending on the sequence composition of the loaded nucleotides. Whether empirically determined via prior sequencing or approximated from a near or distant relative organism, approximate genome length and GC% content can help to achieve the appropriate coverage necessary for acquiring informative NGS data.

Fig. 14.2 Pileup of short reads mapped against the closed reference genome showing an Insertion Sequence of the IS21 family. Blue reads represent those that map to a single genomic location, whereas yellow reads represent those that map to multiple (in this instance 10, Fig. 14.1) separate sites throughout the genome. Extended, repetitive genetic elements are often responsible for termination of contig extension in assemblies based on short reads alone

14.3.2 Coverage

In the context of WGS, coverage generally refers to the mean read depth across a given genome assembly. Calculated by the formula developed by Lander and Waterman, referred to as coverage redundancy, coverage is the product of read length by read count, over the genome size of the organism being sequenced (Eq. 14.1) [11].

$$Coverage = \frac{Read\ length\ (bp) \times Number\ of\ Reads}{Genome\ Size} \tag{14.1}$$

Greater coverage, when equitably distributed across a de novo assembly, can improve de novo assembly quality and promote more accurate variant detection in reference-based assemblies. Depending on platform read length, accuracy, and the nature of the sequenced genome itself, greater coverage can to an extent lead to more complete assemblies composed of longer and less numerous contigs. This can lead to de novo genomic assemblies that are closer representations of complete chromosomes or plasmids. Non-NGS de novo WGS assemblies tend to have coverage in the range of 5–20X, largely due to the costs associated with higher coverage. Modern NGS platforms are, by operator specification, often capable of producing coverage of bacterial genomes in the range of 10^1–10^3X. What constitutes desirable coverage is circumstantial and depends on the goals of the operator. More observations of a given site can increase the confidence of the assembly algorithm in interpreting the validity of a consensus sequence presented by the majority of overlapping reads, provided those reads are uniquely mapping, informative reads [12]. Contig extension, which is handled differently by various assembly programs, requires that the read stem from a template sequence belonging to a sufficiently unique genomic locus. This tends to be an issue with any repetitive nucleotide sequence that exceeds the length of the average read produced by the sequencing platform. Any read lacking some portion of genomic sequence flanking the repetitive sequence is uninformative and often leads to observable coverage spikes in mapped reads against the repetitive region. For purposes of sample multiplexing and efficient

use of often expensive sequencing reagents, coverage should be estimated before sequencing. Additional factors such as sample/library preparation involving a PCR amplification step can contribute to over-representation of select sequences in the sequenced reads, amplicon and shotgun-based libraries may demonstrate different error profiles [13, 14]. The operator can determine an appropriate coverage depth for their sequencing needs. The other variable of the Lander-Waterman equation, read length, tends to be much more sequencing platform-specific. We discuss read length and read accuracy in the context of specific NGS platforms.

14.3.3 Accuracy

Accuracy, the ability to correctly determine the nucleotide present at a given locus, is affected by both the accuracy of the sequencing platform and the actions of the computational assembly process on the sequenced output. In Table 14.1, we provide a brief overview of several modern sequencing platforms frequently used for bacterial WGS. Platforms differ in the types of error profiles they exhibit. Some platforms have variance in read accuracy when challenged with difficult polynucleotides, including those that are compositionally biased (e.g., high/low GC content, homopolymers, highly repetitive DNA). Others are associated with error profiles that are relatively unbiased towards nucleotide composition, but display much lower initial read accuracy and must be compensated through increased coverage. Genomic DNA library preparation is often highly prescribed; routinized protocols ensure, for instance, that the DNA sample that is ultimately input into the platform is the correct length, concentration, single/double stranded, and properly ligated with adaptors. A PCR amplification step may or may not be part of the process of preparing the sample. While beneficial in that PCR can often allow for sequencing from a smaller starting amount of DNA, PCR can also lead to the generation of PCR-induced mutations and the under/over representation of certain sequences based on nucleotide content [15]. This is more likely to be a concern during work with short-read sequencers from the Illumina and Ion torrent platforms, which often utilize a PCR step as part of the library preparation process [14, 16, 17]. Although not necessarily a problem, a PCR step may impact quality of the resulting assembly in certain circumstances. The long-read sequencers from the Pacific Biosciences and Oxford Nanopore platforms tend to have low raw-per-read accuracy relative to the short-read platforms with unbiased and biased errors respectively.

14.4 Short-Read Sequencing

A wide variety of sequencing platforms exist and many determine nucleotide sequence through fundamentally different chemical processes. One way of demarcating platforms is to consider the types of output they provide and how that output relates to the project goals of the user. Several families of sequencing

platforms, notably the Ion torrent and Illumina products, generate short sequenced reads generally in the range of ≤500 bp that tend to assemble into genomic assemblies composed of contigs. Shorter reads are often suitable for de novo assembly of small genomes, and resequencing functions when one has a closed or scaffolded genome and wants to increase coverage over certain regions or across the entire assembly. Bacterial genomes assembled from short reads alone will typically assemble into contigs. Depending on the reason why the contig ended, increased coverage may result in an assembly with fewer and longer contigs, particularly if a region simply was not sufficiently sequenced due to low coverage. But if the contig terminates because of a repetitive DNA sequence that exceeds the average read length, extra coverage depth is unlikely to result in an assembly that further integrates that contig into a longer sequence.

The Ion Torrent platforms (https://www.thermofisher.com) operate via semiconductor chips with emulsion-based, clonally amplified, bead-bound single-stranded nucleotide sequences. The chips are flooded with dNTPs in a set sequence and sensors are positioned to detect hydrogen released from the synthesis reaction that occurs when the cognate base is made available to the template-bound polymerase [18]. In Illumina platforms (https://www.illumina.com), at the start of sequencing, the sample DNA has been fragmented, flanked by adaptor sequences, and exists in single-stranded form. At the start of the sequencing run, an indexed end of the ssDNA fragment becomes bound to the flow cell. This step is followed by local bridge amplification of the DNA fragment, ultimately leading to the creation of a discrete cluster region on the flow cell consisting of many copies of the amplified fragment. The polymerase binds the complement of the adaptor not bound to the flow cell, bases containing labeled fluorophores that block the 3′ hydroxyl groups are added, excited/imaged, and cleaved of the fluorophore, regenerating the 3′ hydroxyl [3]. This process continues for a user specified and kit/platform limited number of cycles.

The applications of reference mapped and de novo genome assemblies produced by short reads are numerous. A closed genome is often unnecessary for identification of the organism at the species rank, and short-read assembles often prove suitable for more complex typing schemes. During work with an assembly derived from a pure-culture isolate, a simple nucleotide BLAST database search using a large contig is often sufficient for a match at the genus/species rank. A local BLAST query against a large curated 16S rRNA dataset such as Silva (https://www.arb-silva.de/) can also be particularly informative. A variety of approaches are potentially available, including single nucleotide polymorphism (SNP)-based trees, core genome multi-lous sequence typing (cgMLST), MLST, and k-mer-based trees. It is not uncommon during an outbreak investigation for a large number of specimens to be collected. The short-read platforms tend to have well-developed options for multiplexing large numbers of isolates on a single sequencing run, allowing for parallel sequencing of isolate's genomes and more efficient use of resources (Table 14.2).

Table 14.2 Platform specifications

Platform	Manufacture	Max read length (bp)	Max reads produced	Max output	WGS applications	Notes	Links
MiniSeq	Illumina	2× 150 bp	25 million	7.5 Gb	Viruses, bacteria, small eukaryotes/targeted sequencing	• High output kit	[1]
NextSeq 550	Illumina	2× 150 bp	800 million	100–120 Gb	Virus, bacteria, eukaryote	• High output kit	[2]
MiSeq	Illumina	2× 300 bp	44–50 million	13.5–5 Gb	Virus, bacteria, small eukaryote/targeted sequencing	• Reagent kit v3	[3]
HiSeq 2500	Illumina	2× 250 bp	4 billion	1000 Gb	Virus, bacteria, eukaryote	• High output mode	[4]
PacBio RS-II	Pacific Biosciences	Max: >60 kb Top 5% >35 kbp	0.365 million	7.6 Gb	virus, bacteria, eukaryote	• Per smart cell, up to 16 per run	[5]
Sequel	Pacific Biosciences	Max: >60 kb Top 5% >35 kbp	0.365 million	7.6 Gb	Virus, bacteria, eukaryote	• Per smart cell, up to 16 per run	[6]
Ion PGM	Ion Torrent	400 bp	4–5.5 million	1.2–2 Gb	Virus, bacteria, small eukaryote/targeted sequencing	• Ion 318 Chip v2 BC	[7]
Ion S5/S5 XL	Ion Torrent	400 bp	60–80 million	10–15 Gb	Virus, bacteria, small eukaryote/targeted sequencing	• Ion 540 chip; Ion 520 and Ion 530 Chips allow for 600 bp reads	[8]
Ion Proton	Ion Torrent	200 bp	60–800 million	10 Gb	Virus, bacteria, small eukaryote/targeted sequencing	• Ion PI chip	[9]
MinION	Oxford Nanopore	>100 kbp	Read length dependent	10–20 Gb	Virus, bacteria, small eukaryote/targeted sequencing	• Per flow cell	[10]

[1] https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/miniseq-reagent-kit.html
[2] https://www.illumina.com/systems/sequencing-platforms/nextseq/specifications.html
[3] https://www.illumina.com/systems/sequencing-platforms/miseq/specifications.html
[4] https://www.illumina.com/systems/sequencing-platforms.html#
[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678779/
[6] http://www.pacb.com/smrt-science/smrt-sequencing/read-lengths/
[7] https://tools.thermofisher.com/content/sfs/brochures/PGM-Specification-Sheet.pdf
[8] https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html
[9] https://tools.thermofisher.com/content/sfs/brochures/CO06326_Proton_Spec_Sheet_FHR.pdf
[10] https://nanoporetech.com/products/minion

14.5 Long-Read Sequencing

One of the drawbacks of short-read sequencing is that de novo assemblies utilizing only short reads rarely assembly into a closed genome for most bacterial and eukaryotic organisms. Closed genome assemblies from short-read data often rely on a reference or additional targeted sequencing methods. The primary reason for this drawback is that if a repetitive region of DNA exceeds the average read length produced by the sequencer, a large quantity of reads will appear identical and the computational assembly program is unable to appropriately place them in the assembly.

There are many situations in which a closed genome is highly desirable for investigating the pathogenicity of an organism. For instance, many virulence factors are horizontally trafficked on composite transposable elements. For instance, it may be of interest to examine the full length of that mobile sequence for unique virulence factors, to determine whether it is on plasmid or chromosomal sequence, or exploring flanking genes for comparative genome analysis. Although increasing coverage depth may be sufficient to ensure the entirety of the sequence assemblies on a single contig, if it is composed of largely non-redundant nucleotides, longer reads are necessary to overcome longer repeat regions. Mobile genetic elements can operate in ways that substantially increase nucleotide sequence redundancy across a genome. Bacterial ISs often fall in the range of 1–2 kbp in length, which is in excess of the average read length produced by most non-paired end short-read sequencers [19]. Some have a copy-paste duplication feature that enables further propagation of the redundant sequence throughout the genome [20]. Through exploitation of endogenous and transferred homologous DNA repair processes, insertional and chromosomal material can be duplicated and altered in numerous ways [21]. As illustrated by Fig. 14.1, examination of the ends of contigs generated from de novo assembly of MiSeq 250-bp paired reads often align with annotated insertional sequence transposase coding sequences that are present at many different sites throughout the bacterial genome. Several sequencing platforms produce reads that average in the range of 5–100 kbp, which is often sufficient to overcome repetitive sequences and produce complete bacterial genomes.

The PacBio RS-II and Sequel platforms can produce a wide range of read lengths depending on user library preparation, with the longest 5% exceeding 35 kbp. As a result, the platform is quite attractive for generating closed bacterial genomes and plasmids. The raw reads have a fairly high single pass error rate of 11–15%, which consists primarily of indel type errors [22, 23]. Each base would have roughly an 11–15% chance of being erroneous if one were only evaluating a single-sequenced read produced by the polymerase. These errors are reported to be unbiased in regards to the nucleotide content. The circular nature of *SMRT-bell* library enables multiple sequencing passes, allowing for generation of subreads that can be utilized for determination of the consensus nucleotide at a given site during genome assembly [22]. In concert with the data acquired from parallel sequencing reactions

overlapping the same site(s), which further contribute to coverage depth, final assemblies attain accuracy in excess of 99.999% [22, 24]. Figure 14.2 shows a close up of MiSeq short reads mapped to the PacBio consensus sequence of Fig. 14.1. The yellow pileup represents non-specific reads that map to the template in multiple locations across the genome. This particular pair of genes encodes an IS21 insertion sequence, which reoccurs seven times with 100% nucleotide identity throughout the genome. Such features often result in contig termination or misassembly in short-read-only assemblies that lack a reference sequence. Long-read sequences are often critical to studies seeking to perform the highest resolution typing possible between closely related bacteria of the same species and those studying broad genomic arrangement.

Although all previously discussed platforms have operated in distinct ways, they all utilize the principle of sequencing by synthesis. The DNA sample of interest is prepared in a way conducive to platform operation, and the platform determines the sequence by synthesizing a complement to the input DNA. The relatively new Oxford Nanopore MinIon platform instead operates by utilizing biological pores and measuring changes in conformation associated with each nucleotide that passes through them, in effect bypassing the need for synthesis on of the input molecule [25]. Error rates in sequenced reads are in the range of 12–35% during a 2-dimensional (2D) double pass sequencing run [26, 27]. MinION reference consortium data from 2017 utilizing the new R9.0 chemistry had a total 2D error rate in reads of 7.5% and were associated with bias towards reads with greater GC content as the run progressed [28]. A reported de novo assembly of a 4.6-Mbp *Escherichia coli* genome with 30× coverage achieved an accuracy of 99.4% [29]. Low accuracy and a developing bioinformatic suite has so far mostly relegated MinION use to reference mapping or the scaffolding of contigs produced by other platforms. This is likely to change as the platform becomes better established and the chemistry continues to be updated. MinION determines base identity as a nucleotide passes through a nanopore and past changes to the pore have had significant impacts on accuracy [28, 30].

The low cost, rapid library preparation, real-time accessible, long reads produced by the MinIon make it a unique addition to the existing field of NGS platforms. In one study, the longer reads were used to scaffold across a repeat region in *Salmonella typhimurium* unable to be resolved using short-read data alone, in turn enabling investigation of an antibiotic resistance island [31]. A hospital-based study tracking a *Salmonella* disease outbreak in multiple patients, using cultured sample derived from patient stool, was able to acquire sufficient data to identify the sample as a *Salmonella* sp. in 20 min, and as serovar Enteritidis within 40 min of sequencing [32]. Several studies utilizing Illumina platform short reads to polish accuracy coupled with long-read scaffolding produced by the MinION were able to assemble a closed bacterial chromosome and a variety of closed plasmids [33, 34].

14.6 Using Next Generation Sequencing Data for Pathogen Detection

Short-read sequencing is generally the starting point for most bacterial sequencing projects. Short-read sequencing platforms tend to allow for the multiplexing of large numbers of isolates at coverage levels acceptable for de novo assembly or read mapping within a single run. Short reads have generally proven more cost effective than their long-read counterparts in terms of both capital and reagent cost. De novo assemblies built from WGS data produced by short-read sequencers are sufficient for purposes of gross identification at the genus/species rank, gene annotation, and investigation of gene products, and can be subjected to a number of traditional and NGS enabled subtyping methods. Long-read sequencers can produce output that can assemble into a closed genome. A closed genome can be used to enable high resolution typing by serving as a reference for very closely related isolates, and allow for the investigation a variety of genomic features and organizations including many associated with pathogenesis and horizontal gene transfer.

This is particularly the case when the underlying genomic diversity in bacteria of a given species is vast or unknown. Botulinum neurotoxin (BoNT), the causative agent of foodborne botulism, is the most potent known biological toxin as estimated by its LD_{50} [35]. BoNT is produced by several species of clostridia including the polyphyletic *C. botulinum* from the *bont* gene, which generally occurs within horizontally trafficked gene clusters roughly 15 kbp in length [36]. Despite most *bont* gene clusters sharing a conserved gene organization/synteny, they often lack broad nucleotide identity. At least eight antigenically distinct BoNT serotypes have been identified and hybrid toxins derived from recombination between serotypes also exist; amino acid identity across serotypes can be as low as ~30% [37–40]. The *bont* gene clusters can be part of chromosomes, plasmids, and phages; bi/tri-toxin producing strains of *C. botulinum* containing multiple *bont* clusters within the same genome have been reported [41]. The relatively limited availability of complete genomes (25 in GenBank as of August 2017), rarity of outbreaks, and shear number of potential combinations of species/strain and toxin serotypes, would generally favor de novo assembly over read mapping when evaluating a newly sequenced isolate or strain genome. Short-read sequencing is generally a strong starting point and with sufficient coverage may result in the inclusion of the entire toxin cluster in a single contig. Long-read sequencing might also be beneficial to ensure the toxin cluster(s) are complete, and can be analyzed with less ambiguity as to their genomic contexts. Additionally, horizontally mobile elements often co-occur at the flanks of the toxin clusters and can frequently terminate contig extension during assembly or result in misassembly. The pathogen genome being sequenced can have unique implications for sequencing, but a de novo assembly produced from short reads is often a strong place to start.

If a wealth of high quality reference genomes already exist, short-read data can be sufficient to infer important genomic organizations with high confidence based on read mapping against a reference sequence. Read mapping is routinely utilized to determine the presence or absence of certain genes and genomic regions reads from

S. enterica relative to reference strains and plasmids [42, 43]. The read-mapped consensus sequence, in the absence of large quantities of unmapped reads and large coverage gaps, can be used to generate a SNP matrix from variant SNPs between the mapped isolate short reads and the closed reference genome. The SNP matrix can be used to calculate distance to generate a phylogeny. Creating a SNP matrix is a complex process and is generally only desirable in situations during which 10s–100s of closely related isolates differ from a closed reference genome at only several hundred SNPs [44]. The resolution provided between strains at this level is unrivaled and allows for extremely detailed subtyping between highly clonal isolates. To explore the relationship between 47 isolates of *Listeria monocytogenes* derived from clinical and food samples in association with a listeriosis outbreak and their relationship to several isolates from a separate outbreak that appeared identical by Pulsed-Field Gel Electrophoresis (PFGE), Chen and colleagues determined closed genome sequences for a pair of isolates to serve as high-quality references and to clarify several putative prophage regions. A SNP-based distance matrix was able to discriminate between the two outbreak strains [45]. When integrated into a database containing a wealth of sequenced clinical, food, and environmental isolate genomes and collection metadata, the subtyping resolution provided by NGS revolutionized outbreak response and outbreak detection. Currently this approach is mostly limited to several heavily sequenced pathogen genomes. However, increasing access to NGS technology, the emergence of a low cost long-read sequencer, and the immense growth in publicly available sequence data may soon make high-resolution subtyping available and attractive to a broader range of investigators.

Bacterial genomes are haploid and should be invariant in regard to nucleotide composition at a given site if the sample was derived from pure culture. Novel mutations may arise during the culturing process and each platform has its own quirks regarding raw read accuracy. Assuming the absence of sample contamination or extremely low coverage, these features are unlikely to interfere with generation of an assembly sufficient for gross identification at the genus/species rank. For certain WGS detection goals, such as the precision subtyping of isolates of the same species through a SNP matrix, it is important to have an awareness of how sample preparation and platform selection can impact accuracy. Short reads can be useful, even when the research goals might strongly suggest the value of obtaining a closed genome. Short-read sequences can be used to reduce, or alter, the error profile present in the closed genome sequence(s). The SPAdes genome assembler (http://cab.spbu.ru/software/spades/) allows hybrid assemblies that can utilize both long and short reads from the platforms surveyed above [46], whereas the program Pilon can use high coverage short reads from a short-read sequencer to detect and correct variants within larger contigs, scaffolded assemblies, or closed bacterial genomes [47].

In terms of software for de novo whole genome assembly, investigators are presented with a wealth of options. We will not delve too deeply in to this topic, but we provide a brief overview of options. Many NGS platforms offer packages that provide options for accessing an in-house or affiliated assembly suite for use with the data created by their sequencers. Consolidated commercial assembly suites such as CLC genomics workbench (QIAGEN Bioinformatics, Denmark) and

Geneious (Biomatters, New Zealand) provide support for a wide range of read types. Additionally, an extremely wide variety of freely available assembly programs, such as Velvet and SPAdes, are available [46, 48]. Of the last category, several options provide graphic user interfaces, but many require use of a command-line and Unix shell. Depending on intended application, it may be of additional benefit to review the manual and/or documentation of the assembly program being used. Assembly programs often operate using either an overlap-layout-consensus or de-bruijn-graph based algorithm which can impact computational performance and assembly quality [49]. While beyond the scope of our review, how the assembly program interacts with factors including the genomic makeup of the sequenced organism and the sequencing platform can significantly impact the completeness and accuracy of the final assembled genome.

Over the past decade, NGS has mostly played a supportive role in pathogen detection. NGS is frequently performed on a microbial culture that has already been selectively isolated and uses either as an additional confirmatory test alongside other tests [e.g., metabolic tests, primer specific polymerase chain reaction (PCR)], or for typing purposes after the identity has been largely confirmed by other means. We will explore how NGS has been integrated into existing typing schemes using Non-typhoidal salmonellae, the leading cause of bacterial foodborne illness, as an illustrative example.

14.7 Applications of Sequencing Data for Pathogen Detection

14.7.1 *Salmonella enterica* in a Public Health Context

Non-typhoidal *Salmonella* are estimated to be responsible for one million annual cases of foodborne illness in the United States and are the leading cause of hospitalization amongst 31 foodborne pathogens [2]. As highly prevalent bacterial pathogens associated with significant morbidity, salmonellae's historic and continuing importance as a pathogen avails the investigator access to a wealth of data concerning the detection, identification, and epidemiological typing of *Salmonella*. In a case of suspected foodborne *Salmonella*, well established, selective isolation protocols from a wide variety of foods are readily available, as are serological tests, metabolic tests, and visual guides for identifying typical/atypical colony morphology [50]. These methods for isolation and confirmation are well established and validated. Despite having undergone revisions over time, changes tend to be incremental resulting in a set of stable, streamlined, and regulatory compliant methods that are accessible by local and clinical health authorities. Although these procedures reliably provide genus confirmation, the selective isolation process alone is generally insufficient for detailed further typing at the species/subspecies ranks. Concerted efforts to surveille *Salmonella* to investigate and prevent their entry into the food supply have been ongoing for 70 years. The ability to categorize and differentiate *Salmonella* strains is central to an effective surveillance program and a multitude of salmonella typing schemes have been used in pursuit of the best means of differentiating between strains (Fig. 14.3).

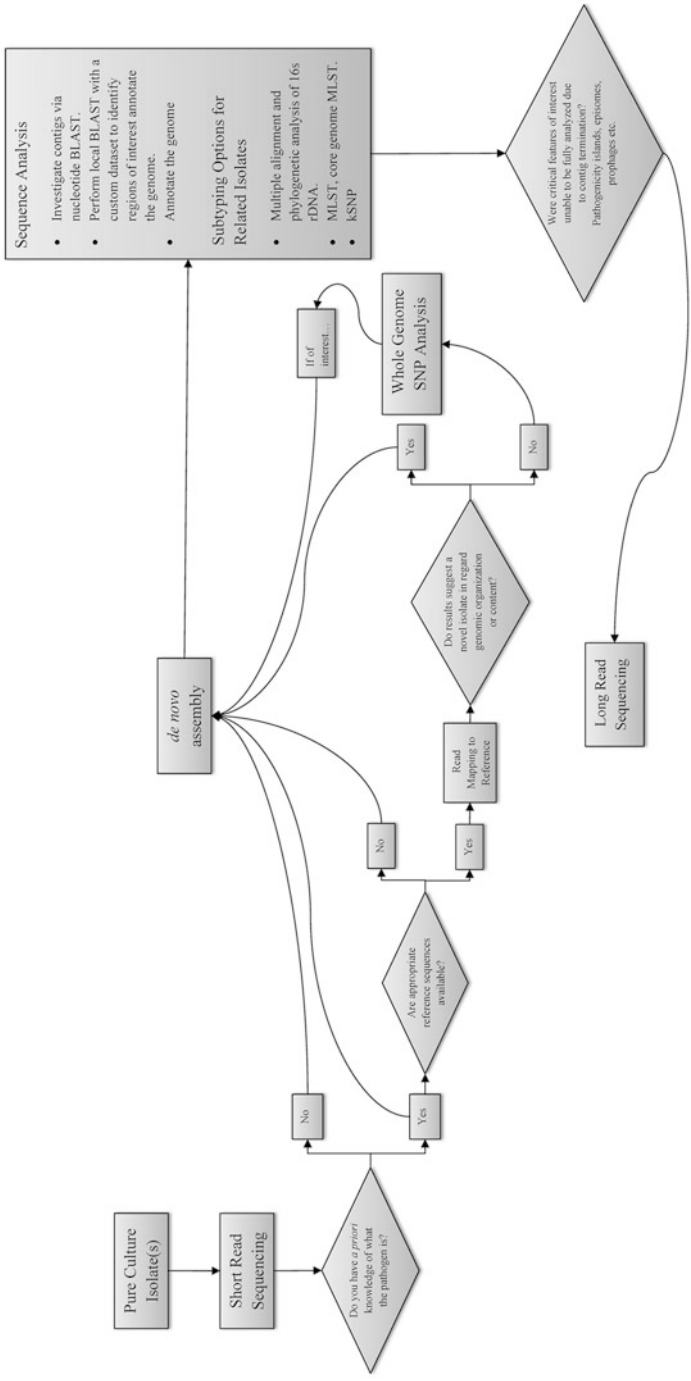

Fig. 14.3 A generalized sequencing scheme for a bacterial pathogen

14.7.2 WGS and Specialized Epidemiological Databases Enable Pathogen Typing at Unparalleled Resolution

Outbreak response is bolstered substantially through the coupling of surveillance metadata with accurately, and specifically typed outbreak isolates. In the U.S., national surveillance efforts for salmonellae were first established in 1963 [51]. Initial surveillance provided valuable insight into the complex ecology of *Salmonella* as pathogens within the food supply ranging from improved understanding of what constituted high risk foods, the impact of mechanized processing activities, and the potential utility of exploiting biochemical differences between strains, trace contaminated foods back to their probable source [52–54]. Utilizing serotyping data, surveillance efforts in the 1960s–1980s were occasionally able to trace food back to their probable source across international borders, though most source tracking successes from this time period occurred at the national level [55, 56]. The increasing globalization of the food supply chain coincided with an increase in the frequency and scale of foodborne salmonellosis outbreaks [57]. Greater diversity in the supply chain and increased opportunities for exchanging products and pathogens placed further emphasis on the importance of being able to discriminate between salmonellosis outbreak strains.

14.7.3 Salmonella Serotyping and Inferring Salmonella Serotyping via PCR

Salmonella classification/typing by serotype under the White-Kauffmann-Le Minor scheme organizes *Salmonella* strains by surface antigen [58–60]. The method predates DNA based molecular methods, remains in practice today, and illustrates how advances in molecular typing methodologies can build upon one another. As of the last major WHO update in 2007, 2557 *S. enterica* serovars were described based on a combination of O, H, and Vi antigens [59]. The subtyping process takes several days, is labor intensive, and often proves impractical to perform on more than a small fraction of individual isolates associated with a given disease outbreak.

A variety of approaches, including specialized high-throughput mechanical/automated serotyping, PCR-based methods, and most recently NGS have been used to decrease the time and labor necessary to determine serotype [61–63]. Allele-specific PCR methods were developed that target the genes encoding the antigenic determinants enabling inference of serotypes from analysis of PCR amplicon band patterns on electrophoretic gel. Such methods are reported to have reduced the time from start to finish to 5 h with high concordance between physical typing and the PCR method (108 of 111 tested isolates in one such study) [61]. Comprised of the entire genomic nucleotide content of a given organism or even a given cell, WGS data provides unparalleled insight into the gene content, organization, and predicted coding sequences present. Most allelic typing schemes are or can be made compatible with WGS data.

14.7.4 Inferring Salmonella Serotype via WGS Assembly

The Seroseq webserver accepts WGS reads in fastq file format and assigns *Salmonella* serotypes by searching for the specific alleles associated with the *fliC*, *fljB*, and *rfb* genes that encode the serovar-determinative proteins [63]. NGS allows the preparation and multiplexed whole genome sequencing of 100s of individual *Salmonella* isolates at once. Although the preparation time involved for WGS on a short-read sequencer with high capacity for multiplexing, such as the Illumina benchtop platform family, for 100–200 samples is generally several days, it can be conducted by a single individual using a streamlined process that does not require the use of primer-specific PCR in conjunction with gel electrophoresis to visualize results. With WGS involved as a routine part of outbreak response, this approach allows an assessment of serotype from many samples that would not have been subjected to serological analysis. This inclusion also enables generation of trees that reflect antigenic, rather than phylogenetic ancestral traits, which enable tracking horizontally mobilized genes including those that confer antibiotic resistant and virulence traits.

14.7.5 NGS Data Can Often Be Integrated with Existing Allelic PCR Based Typing Schemes

Most short-read assemblies of salmonella isolates are equally amenable to MLST, cgMLST, and any sort of allele-based typing scheme that the user desires. Some have argued for the replacement of *Salmonella* serotyping with an MLST scheme composed of allelic variants of several well conserved housekeeping genes that together provide a level of discriminatory capacity. This method is generally more consistent with phylogenetic trends observed when using higher resolution methods [64]. The value of WGS data is that they allow the investigation of both the phylogenetic ancestry of the organism and the phylogeny of any subset of nucleic acids common to the isolates being investigated. The National Antimicrobial Resistance Monitoring System (NARMS) characterizes and monitors spread of antimicrobial resistance (AMR) genes. Investigators were able to compare rapidly the antimicrobial resistance profile of 640 *Salmonella* isolates by searching against several AMR databases with 90% correlation between genotype and in-vitro tested phenotype [65]. Specialized datasets such as those specializing in AMR grow in value as more functionally characterized genes/protein-encoding sequences are entered into them. In such a setting, discordance between phenotype and genotype can lead to novel discoveries and flag flaws in predictive workflows. Furthermore, this setting enables the utilization of sequence data in conjunction with associated gene products and other publicly available sequence data that can provide critical insight into pathogenesis at the molecular level.

14.7.6 WGS Is an Unparalleled Tool for High Resolution Bacterial Subtyping

Although collection of metadata is an invaluable tool in source tracking, it is only as useful as the resolving power of the typing method used. Pulsed-Field Gel Electrophoresis (PFGE), a method developed to separate large nucleic acids, typically utilizes a restriction digest that generates a unique pattern on an electrophoretic gel and often comparison of the pattern is sufficient to distinguish between different *Salmonella* strains [60]. In the U.S., PFGE has been heavily utilized for *Salmonella* typing as part of the PulseNet surveillance network and within 5 years of operation, PulseNet USA contained 110,000 PFGE *Salmonella* profiles [66, 67]. Although generally able to discriminate between *Salmonella* strains, PFGE can be insufficient when very closely related strains are analyzed and lacks the resolution necessary to indicate differences between isolates from the same outbreak [68, 69]. To further bolster its discriminatory capacity, PulseNet incorporated additional amplicon based methods for *Salmonella*. One such typing scheme, multi-locus variable number tandem repeat analysis (MLVA), is PCR = based and utilizes predesigned primers that produce amplicons that vary in size if variation in repeat length is present across strains. Due to the high concordance between MLVA and PFGE results and the increased resolution of MLVA over PFGE, MLVA may be used following PFGE to increase resolution between visually non-discernable strain pulsed-gel profiles [69].

NGS both bolsters and disrupts previous typing schemes for salmonellae. NGS enables the rapid whole genome sequencing (WGS) of *Salmonella*, and virtually any culturable bacterial organism. One can directly compare two or more *Salmonella* strains/isolates across all available sites, and this is readily achieved through a whole genome SNP analysis using high quality reference strains. For typing purposes, the resolution of WGS data alone is theoretically equal to the total nucleotide count present in the genome assembly, including the chromosome and the sum of any present plasmids. *Salmonella* genomes typically range in length from 4.4 to 5.8 Mbp [70]. The data produced through cross-genome analysis of WGS data allow for greater resolution than previously discussed methods, and can differentiate between *Salmonella* strains that cannot be distinguished by PFGE. An early application of a WGS as a subtyping tool was demonstrated through its ability to distinguish *Salmonella* Montevideo isolates, which generally appear identical in PFGE. WGS of 47 strains of *Salmonella* Montevideo led to the identification of 23 informative SNPs that concurred with their respective disease outbreaks upon analysis within a derived phylogenetic tree [68].

As of 2018, 5 years after its creation, the GenomeTrakr database contained just over 105,000 genomes of *S. enterica* isolates. Government and academic laboratories now routinely use WGS as part of their pathogen detection workflows and the scale of this cooperation and the accessibility of the data continue to grow. The pathogen detection (https://www.ncbi.nlm.nih.gov/pathogens/) data are further enhanced by a wealth of metadata that enables rapid outbreak response and new avenues for outbreak prevention [71]. Greater resolution allows for more precise matching between environmental, food, and clinical samples.

14.7.7 WGS, Public Databases, and Bioinformatics Enable New Means for Discovering Putative Virulence Factors

Returning to the example of foodborne botulism, the preformed BoNT is the causative agent and detection protocols generally focus on identifying the BoNT serotype as it is often the primary indicator of disease phenotype, duration and progress [72]. Contrary to salmonellosis, foodborne botulism detection methods generally prioritize toxin detection and serological typing first, with a secondary emphasis on isolating and growing the organism. WGS has provided insight into the horizontal transfer, recombination, and evolution of BoNTs. Over the past decade alone, NGS has led to the discovery of new hybrid toxins, silent toxins, novel serotypes, and the discovery of BoNT-like toxin homologs in organisms beyond members of the genus *Clostridium*. In doing so, NGS has forced a spirited debate over the very definition of a BoNT serotype and demonstrated how the targeted sequencing of known pathogens supports this work.

14.7.8 Role of WGS in Investigating a Hybrid Botulinum Toxin Consisting of Two Serotypes

In 2013, investigators described a novel botulinum toxin produced by a bivalent Group I *C. botulinum* strain isolated from infant stool in a case of infant botulism. Initial efforts to serologically type the toxins using Centers for Disease Control and Prevention (CDC)-provided monovalent polyclonal antibodies for known serotypes A–G resulted in identification of a B serotype toxin and an unknown toxin non-neutralized by the antibodies referred to as BoNT/H [37]. WGS of the clostridium strain revealed that the coding sequence for the unknown BoNT serotype had a mosaic/hybrid composition similar to that of BoNT/F and BoNT/A serotypes. Subsequent investigation indicated the BoNT/H type toxin can be neutralized with BoNT/A antitoxin [38, 39]. Debate persists over nomenclature, the use of research vs. non-research antitoxins in achieving neutralization, antitoxin dosage and potency, and whether the BoNT/FA(H) hybrid represents a novel serotype in its own right [73]. However, determination of the BoNT/FA(H) hybrid nature via WGS meaningfully informed neutralization research that could augment treatment of BoNT/FA(H) intoxination or intoxination by closely related homologs should they emerge again in the future.

BoNT's are significantly diverse across serotypes at the primary amino acid level but substantially conserved in regard to core motifs and domains. Naturally occurring recombinants such as BoNT FA(H) demonstrate that hybrid serotypes can be capable of causing disease. Some serotypes, including BoNT/C, BoNT/D, and BoNT/G, have rarely, if ever been associated with a human case of botulism, though BoNT/C/D and their hybrids are major sources of botulism in wild animals and livestock [74–77]. WGS has revealed that such multivalent clostridium strains such as that producing BoNT/FA(H) and BoNT/B are not particularly unusual, and B serotype clusters with nonsense mutations have also been observed in a number of

isolates [78]. While not human disease causing, toxin fragments and non-implicated serotypes may remain relevant through their capacity to recombine with serotypes more frequently responsible for botulism outbreaks. Heptavalent botulinum antitoxin (HBAT), an equine polyclonal treatment against BoNT serotypes A–G, is the primary treatment for foodborne botulism in the US [79]. Assuming hybrid toxins remain antigenically similar enough to their constituent serotypes, existing polyclonal antitoxin treatments should have the capacity to bind novel hybrid types. HBAT is effective against BoNT/FA(H) [80]. However, WGS has also revealed that additional antigenically distinct serotypes exist. Bioinformatics analysis of WGS data revealed in 2017 the first novel BoNT serotype discovered since 1970 [40, 81].

14.7.9 Discovery of a Novel Botulinum Neurotoxin Serotype via Nucleotide/Protein Databases

A novel serotype, BoNT/X was recently described and represents the first to be identified through bioinformatics. *C. botulinum* strain 111 was initially isolated from a case of infant botulism in 1996 and was observed, at the time, to test positive for BoNT/B [40, 82]. A closed genome assembly of *C. botulinum* strain 111 was released publically by NCBI in 2015 with a plasmid-borne BoNT/B2 toxin gene cluster and a previously undiscovered chromosomal BoNT toxin gene cluster containing a putative BoNT with low identity but broad homology to all known serotypes [AP014696.1]. Experiments utilizing partial and sortase linked recombinant BoNT/X found that it cleaves several traditional and non-traditional SNARE substrates, is antigenically distinct from known serotypes, and causes flaccid paralysis in laboratory mice [40]. It remains unknown whether BoNT/X is expressed by *C. botulinum* strain 111. The ongoing investigation into BoNT/X highlights how bioinformatics-driven investigation of WGS data can allow detection and subsequent analysis of cryptic virulence factors that may not be detectable through other means. BoNT/X evaded detection in the 1990s and was captured through WGS conducted 20 years later. Within 2 years of the sequence being made public, researchers had bioinformatically characterized, artificially synthesized, recombinantly expressed, and demonstrated the enzymatic functionality of a novel BoNT serotype.

14.7.10 Identification of Botulinum Neurotoxin-Like Proteins in Bacteria Beyond Genus *Clostridium*

Interestingly, both the BoNT FA(H) and BoNT X toxins were discovered in dual toxin-producing strains of *C. botulinum*. The *bont* gene cluster exists within flanking ISs, prophages, and plasmids suggesting some degree of horizontal mobility. A complete toxin cluster encoding an enzymatically functional botulinum toxin was recently identified on a plasmid of an *Enterococcus faecium* isolate that was obtained from cattle feces [83]. Like BoNT/X, this cryptic toxin was identified through bioinformatics driven investigation of sequence databases. Even more

divergent BoNT-like-encoding sequences with broad homology to those encoding BoNT have been identified in a range of bacteria [84–86]. Although some of these organisms, including *Enterococcus* and *Weissella*, share a similar ecological niche as *Clostridium* and require similar strict anaerobic conditions for growth, others do not. The sudden marked growth in the apparent prevalence and diversity of botulinum toxin homologs is unlikely to be an isolated occurrence. The increased diversity and depth of sequence databases is revealing that many important virulence factors may have similarly complex horizontal distributions.

Understanding the horizontal transfer of critical virulence factors, coupled with a large global database of sequence data, can provide new insight into virulence factor evolution and potentially flag emerging or previously unknown pathogens for further investigation. In the case of botulinum neurotoxin, this coupling has led to the discovery of distant toxin homologs in bacteria of unrelated species. WGS opens new avenues of cross-species investigation into horizontally transmitted genes and gene products that include many fundamental molecular building blocks that enable bacterial pathogens to cause disease. WGS allows not only the investigation and comparative analysis of the nucleotides present within the sequenced isolate genome, but also the investigation of its gene products, and the utilization of any associated metadata. The integration of these data in the fields of outbreak detection, response, and prevention have enabled the development of robust public health programs.

14.8 Conclusions

NGS has become a technology central to pathogen detection and characterization. WGS is a useful tool for high resolution typing of closely and distantly related bacterial pathogens, and has already changed the landscape of disease surveillance. In contrast to the previous high-resolution typing standard, PFGE, WGS also produces an abundance of data at the gene/allele level that can provide in-silico backwards compatibility with PCR-based typing methods such as MLST. For closely related isolates for which suitable reference sequences exist, a variety of SNP-matrix based typing schemes can be used and for those lacking such reference sequences, a cgMLST approach can provide excellent resolution. In addition, continued advances in NGS technology are beginning to provide a growing number of investigators access to the long reads necessary to generate high quality reference genomes. Data begets data and give rise to new applications of those data. The growth of public sequence databases and specialized pathogen tracking databases with detailed collection metadata inspire myriad new research activities into pathogen behavior, ecology, and evolution.

Over the past decade NGS enabled WGS has rapidly progressed from an experimental technology to a core technology for disease surveillance, response, and prevention. Exciting new applications of NGS technology beyond WGS are now being explored for their potential to augment outbreak response. Sophisticated metagenomic approaches are being increasingly explored as a potential means of

direct detection of pathogens in suspected contaminated substrate with the potential to speedup or bypass the culturing process. RNA-sequencing and ribosomal profiling allow for genome-wide investigation of transcriptional and translational activity which may eventually enable large-scale quantitative analysis of virulence factor expression across outbreak isolates. WGS has invigorated new research on pathogenicity and the future will hold new applications that will arise from this dynamic new technology. Sequencing technology continues to improve at a rapid rate. As the field continues to mature one can also expect to see sequencers emerge that can work with smaller inputs and provide actionable data with quicker turn around and reduced costs.

References

1. Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM. Foodborne illness acquired in the United States—unspecified agents. Emerg Infect Dis. 2011a;17:16–22. https://doi.org/10.3201/eid1701.091101p2.
2. Scallan E, et al. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis. 2011b;17:7–15. https://doi.org/10.3201/eid1701.P11101.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.
4. Fleischmann RD, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;269:496–512.
5. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24:133–41. https://doi.org/10.1016/j.tig.2007.12.007.
6. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135. https://doi.org/10.1038/nbt1486.
7. Rasko DA, et al. *Bacillus anthracis* comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci USA. 2011;108:5027–32. https://doi.org/10.1073/pnas.1016657108.
8. Ravel J, et al. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor". J Bacteriol. 2009;191:445–6. https://doi.org/10.1128/JB.01347-08.
9. Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. Evidence for DNA loss as a determinant of genome size. Science. 2000;287:1060–2.
10. Nowrousian M. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell. 2010;9:1300–10. https://doi.org/10.1128/EC.00123-10.
11. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988;2:231–9.
12. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15:121–32. https://doi.org/10.1038/nrg3642.
13. Aird D, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12:R18. https://doi.org/10.1186/gb-2011-12-2-r18.
14. Ross MG, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14:R51. https://doi.org/10.1186/gb-2013-14-5-r51.
15. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet. 2010;11:31–46. https://doi.org/10.1038/nrg2626.

16. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. Brief Bioinform. 2016;17:154–79. https://doi.org/10.1093/bib/bbv029.
17. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015;43:e37. https://doi.org/10.1093/nar/gku1341.
18. Merriman B, Ion Torrent R, Team D, Rothberg JM. Progress in ion torrent semiconductor chip based sequencing. Electrophoresis. 2012;33:3397–417. https://doi.org/10.1002/elps.201200424.
19. Mahillon J, Léonard C, Chandler M. IS elements as constituents of bacterial genomes. Res Microbiol. 1999;150:675–87. https://doi.org/10.1016/S0923-2508(99)00124-2.
20. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38:865–91. https://doi.org/10.1111/1574-6976.12067.
21. Darmon E, Leach DR. Bacterial genome instability. Microbiol Mol Biol Rev. 2014;78:1–39. https://doi.org/10.1128/MMBR.00035-13.
22. Korlach J (2013) Understanding accuracy in SMRT® sequencing. https://www.mscience.com.au/upload/pages/pacbioaccuracy/perspective-understanding-accuracy-in-smrt-sequencing.pdf. Accessed 10 Jan 2018
23. Rhoads A, Au KF. PacBio sequencing and its applications. GPB. 2015;13:278–89. https://doi.org/10.1016/j.gpb.2015.08.002.
24. Quail MA, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific biosciences and illumina MiSeq sequencers. BMC Genomics. 2012;13:341. https://doi.org/10.1186/1471-2164-13-341.
25. Mikheyev AS, Tin MM. A first look at the Oxford nanopore MinION sequencer. Mol Ecol Resour. 2014;14:1097–102. https://doi.org/10.1111/1755-0998.12324.
26. Ip CLC, et al. MinION analysis and reference consortium: phase 1 data release and analysis. F1000Res. 2015;4:1075. https://doi.org/10.12688/f1000research.7201.1.
27. Lu H, Giordano F, Ning Z. Oxford nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinform. 2016;14:265–79. https://doi.org/10.1016/j.gpb.2016.05.004.
28. Jain M, et al. MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. F1000Res. 2017;6:760. https://doi.org/10.12688/f1000research.11354.1.
29. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12:733–5. https://doi.org/10.1038/nmeth.3444.
30. Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep. 2015;5:11996. https://doi.org/10.1038/srep11996.
31. Ashton PM, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2015;33:296–300. https://doi.org/10.1038/nbt.3103.
32. Quick J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biol. 2015;16:114. https://doi.org/10.1186/s13059-015-0677-2.
33. Risse J, Thomson M, Patrick S, Blakely G, Koutsovoulos G, Blaxter M, Watson M. A single chromosome assembly of bacteroides fragilis strain BE1 from Illumina and MinION nanopore sequencing data. Gigascience. 2015;4:60. https://doi.org/10.1186/s13742-015-0101-6.
34. Wick RR, Judd LM, Gorrie CL, Holt KE. Completing bacterial genome assemblies with multiplex MinION sequencing. Microbial Genomics. 2017;3:1–7. https://doi.org/10.1099/mgen.0.000132.
35. Gill DM. Bacterial toxins: a table of lethal amounts. Microbiol Rev. 1982;46:86–94.
36. Hill KK, Smith TJ. Genetic diversity within *Clostridium botulinum* serotypes, botulinum neurotoxin gene clusters and toxin subtypes. Curr Top Microbiol Immunol. 2013;364:1–20. https://doi.org/10.1007/978-3-642-33570-9_1.

37. Dover N, Barash JR, Hill KK, Xie G, Arnon SS. Molecular characterization of a novel botulinum neurotoxin type H gene. J Infect Dis. 2014;209:192–202. https://doi.org/10.1093/infdis/jit450.
38. Gonzalez-Escalona N, et al. Draft genome sequence of bivalent *Clostridium botulinum* strain IBCA10-7060, encoding botulinum neurotoxin B and a new FA mosaic type. Genome Announc. 2014;2:e01275–14. https://doi.org/10.1128/genomeA.01275-14.
39. Maslanka SE, et al. A novel botulinum neurotoxin, previously reported as serotype H, has a hybrid-like structure with regions of similarity to the structures of serotypes A and F and is neutralized with serotype A antitoxin. J Infect Dis. 2016;213:379–85. https://doi.org/10.1093/infdis/jiv327.
40. Zhang S, et al. Identification and characterization of a novel botulinum neurotoxin. Nat Commun. 2017;8:14130. https://doi.org/10.1038/ncomms14130.
41. Dover N, Barash JR, Hill KK, Davenport KW, Teshima H, Xie G, Arnon SS. *Clostridium botulinum* strain Af84 contains three neurotoxin gene clusters: bont/A2, bont/F4 and bont/F5. PLoS One. 2013;8:e61205. https://doi.org/10.1371/journal.pone.0061205.
42. Bachmann NL, Petty NK, Ben Zakour NL, Szubert JM, Savill J, Beatson SA. Genome analysis and CRISPR typing of *Salmonella enterica* serovar Virchow. BMC Genomics. 2014;15:389. https://doi.org/10.1186/1471-2164-15-389.
43. Wilson MR, et al. Whole genome DNA sequence analysis of *Salmonella* subspecies *enterica* serotype Tennessee obtained from related peanut butter foodborne outbreaks. PLoS One. 2016;11:e0146929. https://doi.org/10.1371/journal.pone.0146929.
44. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Comput Sci. 2015;1:e20.
45. Chen Y, et al. *Listeria monocytogenes* in stone fruits linked to a multistate outbreak: enumeration of cells and whole-genome sequencing. Appl Environ Microbiol. 2016;82:7030–40. https://doi.org/10.1128/AEM.01486-16.
46. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.
47. Walker BJ, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963. https://doi.org/10.1371/journal.pone.0112963.
48. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. https://doi.org/10.1101/gr.074492.107.
49. Li Z, et al. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. Brief Funct Genomics. 2012;11:25–37.
50. Food and Drug Administration (2018) Bacteriological analytical manual. Chapter 5: *Salmonella*. https://www.fda.gov/food/foodscienceresearch/laboratorymethods/ucm070149.htm. Accessed 12 Jan 2018
51. Schroeder SA, Aserkoff B, Brachman PS. Epidemic salmonellosis in hospitals and institutions: a five-year review. N Engl J Med. 1968;279:674–8. https://doi.org/10.1056/NEJM196809262791303.
52. Cohen ML, Blake PA. Trends in foodborne salmonellosis outbreaks: 1963–1975. J Food Prot. 1977;40:798–800. https://doi.org/10.4315/0362-028X-40.11.798.
53. Martin WJ, Ewing WH. Prevalence of serotypes of *Salmonella*. Appl Microbiol. 1969;17:111–7.
54. Wilder AN, MacCready RA. Isolation of *Salmonella* from poultry: poultry products and poultry processing plants in Massachusetts. N Engl J Med. 1966;274:1453–60. https://doi.org/10.1056/NEJM196606302742601.
55. Clark GM, Kaufmann AF, Gangarosa EJ, Thompson MA. Epidemiology of an international outbreak of *Salmonella* Agona. Lancet. 1973;2:490–3. https://doi.org/10.1016/S0140-6736(73)92082-5.

56. Craven P, et al. International outbreak of Salmonella eastbourne infection traced to contaminated chocolate. Lancet. 1975;305:788–92.
57. Rodrigue DC, Tauxe RV, Rowe B. International increase in Salmonella enteritidis: a new pandemic? Epidemiol Infect. 1990;105:21–7.
58. Brenner FW, Villar RG, Angulo FJ, Tauxe R, Swaminathan B. Salmonella nomenclature. J Clin Microbiol. 2000;38:2465–7.
59. Grimont PA, Weill F-X. Antigenic formulae of the *Salmonella* serovars. 9th ed. Paris: WHO Collaborating Centre for Reference and Research on Salmonella; 2007. p. 1–166.
60. Steve Yan S, Pendrak ML, Abela-Ridder B, Punderson JW, Fedorko DP, Foley SL. An overview of *Salmonella* typing. Clin Appl Immunol Rev. 2004;4:189–204. https://doi.org/10.1016/j.cair.2003.11.002.
61. Kim S, Frye JG, Hu J, Fedorka-Cray PJ, Gautom R, Boyle DS. Multiplex PCR-based method for identification of common clinical serotypes of *Salmonella enterica* subsp. *enterica*. J Clin Microbiol. 2006;44:3608–15. https://doi.org/10.1128/JCM.00701-06.
62. Shipp CR, Rowe B. A mechanised microtechnique for salmonella serotyping. J Clin Pathol. 1980;33:595–7.
63. Zhang S, et al. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015;53:1685–92. https://doi.org/10.1128/JCM.00323-15.
64. Achtman M, et al. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. PLoS Pathog. 2012;8:e1002776. https://doi.org/10.1371/journal.ppat.1002776.
65. McDermott PF, et al. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal Salmonella. Antimicrob Agents Chemother. 2016;60:5515–20. https://doi.org/10.1128/AAC.01030-16.
66. Gerner-Smidt P, et al. PulseNet USA: a five-year update. Foodborne Pathog Dis. 2006;3:9–19. https://doi.org/10.1089/fpd.2006.3.9.
67. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, Force CDCPT. PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. Emerg Infect Dis. 2001;7:382–9. https://doi.org/10.3201/eid0703.010303.
68. Allard MW, et al. High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics. 2012;13:32. https://doi.org/10.1186/1471-2164-13-32.
69. Boxrud D, Pederson-Gulrud K, Wotton J, Medus C, Lyszkowicz E, Besser J, Bartkus JM. Comparison of multiple-locus variable-number tandem repeat analysis, pulsed-field gel electrophoresis, and phage typing for subtype analysis of *Salmonella enterica* serotype Enteritidis. J Clin Microbiol. 2007;45:536–43. https://doi.org/10.1128/JCM.01595-06.
70. Land M, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15:141–61. https://doi.org/10.1007/s10142-015-0433-4.
71. Stevens EL, Timme R, Brown EW, Allard MW, Strain E, Bunning K, Musser S. The public health impact of a publically available, environmental database of microbial genomes. Front Microbiol. 2017;8:808. https://doi.org/10.3389/fmicb.2017.00808.
72. Food and Drug Administration (2017) Bacteriological analytical manual. Chapter 17: *Clostridium botulinum*. https://www.fda.gov/Food/FoodScienceResearch/LaboratoryMethods/ucm070879.htm. Accessed 20 Jan 2018
73. Fan Y, Barash JR, Lou J, Conrad F, Marks JD, Arnon SS. Immunological characterization and neutralizing ability of monoclonal antibodies directed against botulinum neurotoxin type H. J Infect Dis. 2016;213:1606–14. https://doi.org/10.1093/infdis/jiv770.
74. Collins MD, East AK. Phylogeny and taxonomy of the food-borne pathogen *Clostridium botulinum* and its neurotoxins. J Appl Microbiol. 1998;84:5–17. https://doi.org/10.1046/j.1365-2672.1997.00313.x.
75. Oguma K, et al. Infant botulism due to *Clostridium botulinum* type C toxin. Lancet. 1990;336:1449–50.

76. Sonnabend O, Sonnabend W, Heinzle R, Sigrist T, Dirnhofer R, Krech U. Isolation of *Clostridium botulinum* type G and identification of type G botulinal toxin in humans: report of five sudden unexpected deaths. J Infect Dis. 1981;143:22–7.
77. Takeda M, Tsukamoto K, Kohda T, Matsui M, Mukamoto M, Kozaki S. Characterization of the neurotoxin produced by isolates associated with avian botulism. Avian Dis. 2005;49:376–81. https://doi.org/10.1637/7347-022305R1.1.
78. Smith TJ, Hill KK, Raphael BH. Historical and current perspectives on *Clostridium botulinum* diversity. Res Microbiol. 2015;166:290–302. https://doi.org/10.1016/j.resmic.2014.09.007.
79. Centers for Disease Control and Prevention. Investigational heptavalent botulinum antitoxin (HBAT) to replace licensed botulinum antitoxin AB and investigational botulinum antitoxin E. Morb Mortal Wkly Rep. 2010;59:299.
80. Pellett S, et al. Purification and characterization of botulinum neurotoxin FA from a genetically modified *Clostridium botulinum* strain. mSphere. 2016;1:e00100–15. https://doi.org/10.1128/mSphere.00100-15.
81. Gimenez D, Ciccarelli A. Another type of *Clostridium botulinum*. Zentralbl Bakteriol Parasitenkd Infekt Hyg Abt I (Orig). 1970;215:221–4.
82. Kakinuma H, Maruyama H, Takahashi H, Yamakawa K, Nakamura S. The first case of type B infant botulism in Japan. Acta Paediatr Jpn. 1996;38:541–3.
83. Zhang S, et al. Identification of a botulinum neurotoxin-like toxin in a commensal strain of *Enterococcus faecium*. Cell Host Microbe. 2018;23(2):169–76.
84. Mansfield MJ, Adams JB, Doxey AC. Botulinum neurotoxin homologs in non-Clostridium species. FEBS Lett. 2015;589:342–8. https://doi.org/10.1016/j.febslet.2014.12.018.
85. Mansfield MJ, Wentz TG, Zhang S, Lee EJ, Dong M, Sharma SK, Doxey AC. Newly identified relatives of botulinum neurotoxins shed light on their molecular evolution. bioRxiv. 2017. https://doi.org/10.1101/220806
86. Wentz TG, et al. Closed genome sequence of *Chryseobacterium piperi* strain CTM(T)/ATCC BAA-1782, a gram-negative bacterium with clostridial neurotoxin-like coding sequences. Genome Announc. 2017;5. https://doi.org/10.1128/genomeA.01296-17