MiSeq® Efficiently Performs Bacterial Typing

A retrospective analysis of CDC samples from the 2011 *Listeria* outbreak with Illumina systems.

Introduction

Listeria monocytogenes was named after Joseph Lister, a pioneer in bacteriology and the father of modern antisepsis. First identified in 1926 as a human pathogen, it was not until the early 1980s that *L. monocytogenes* was recognized as a major source of foodborne illness.

One of the most deadly outbreaks in the United States occurred in 2011, when *Listeria*-contaminated cantaloupes infected 146 people, killing 30[1]. The outbreak spread quickly and was tracked by PulseNet, the U.S. Centers for Disease Control and Prevention's (CDC) national molecular subtyping surveillance system of foodborne pathogens. The outbreak spread to 28 states before it was contained. The PulseNet team at the CDC, and at federal food regulatory agencies and public health laboratories across the country, used pulse field gel electrophoresis (PFGE) to subtype *Listeria* isolates from human cases and cantaloupe samples and track the outbreak, pinpointing the source as a cantaloupe packaging plant in northeastern Colorado.

Even though PulseNet is an effective laboratory-based surveillance system for foodborne bacterial pathogens, one could ask, "Is there a more efficient way of performing bacterial typing?" Public health agencies like the CDC are investigating the value of next-generation sequencing (NGS), and how it can support and enhance the rapid identification and investigation of clusters of foodborne illness caused by foodborne bacterial pathogens such as *L. monocytogenes*.

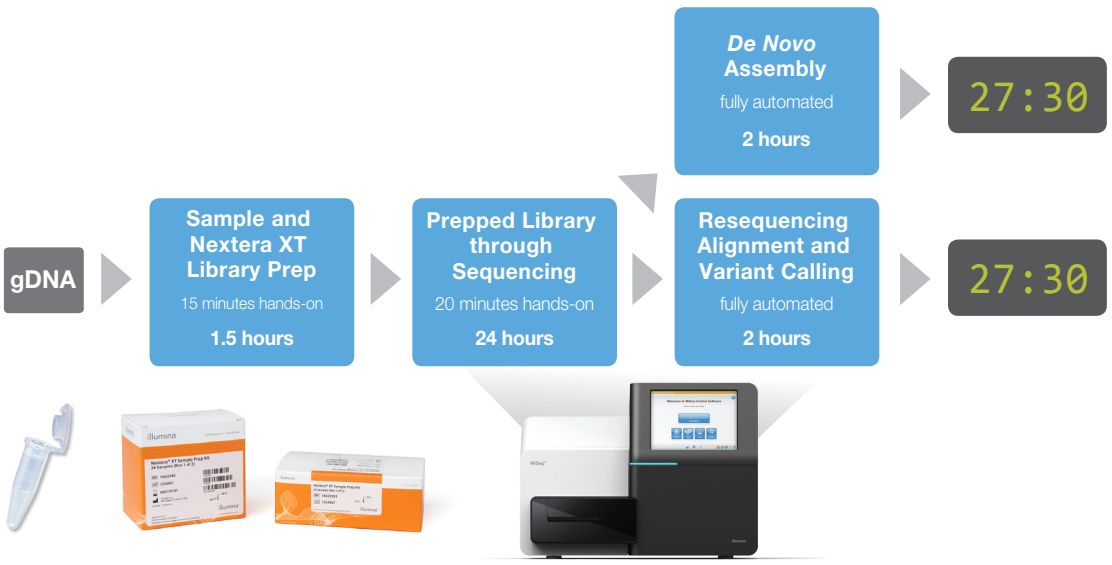
Next-Generation Sequencing Benefits

PFGE is currently the gold standard in epidemiological studies of pathogenic foodborne bacteria, yet the 30–year old technology has limitations. The CDC relies on alternative typing methods when PFGE does not provide definitive results. For example, in cases where:

- Results are difficult to interpret due to the ambiguity of same-size fragments representing different parts of the genome;
- There is a need to further discern common patterns; and
- When bacterial strains are untypable

Bacterial typing using NGS is an emerging alternative to PFGE, with instruments such as the Illumina high-throughput HiSeq® and lower-throughput MiSeq® systems offering simpler workflows, higher resolution, and a universally applicable bacterial subtyping method. These systems are particularly well suited to operate within the PulseNet's "hub and spoke" network, where high sample volumes are processed at CDC in the PulseNet central laboratory, with lower sample volumes analyzed in regional laboratories.

Figure 1: MiSeq Sequencing Workflow



The MiSeq system delivers results in hours rather than days through a streamlined workflow and fast Nextera® sample preparation technology. Times shown reflect specifications for Nextera XT Library Prep and MiSeq system as of August 2012.

Table 1: MiSeq and Genome Analyzer Run and Assembly Metrics

	Sample #1		Sample #2	
	Genome Analyzer IIx	MiSeq	Genome Analyzer IIx	MiSeq
Number of reads	4,386,317 (used: 500,252)	519,260	4,971,395 (used: 675,406)	622,190
% Bases ≥ Q30	84.03%	90.68%	83.71%	91.19%
Assembly Size	2,872,167	2,873,471	2,909,352	2,896,691
Number of Contigs	442	128	248	158
N50 [bp (Contigs)]	13,920 (61)	55,000 (15)	30,716 (28)	36,962 (24)
Average Contig Size (bp)	6,498.11	22,448.99	11,731.25	18,333.48
% Gaps in Scaffolds	0.0003	0.03	0.002	0.02
Max Contig (bp)	73,736	235,782	106,096	239,135
% GC	37.88	37.89	37.85	37.92

NGS is not new to the public health community, with a growing number of applications demonstrating the utility of NGS in clinical microbiology. Public health laboratories in Canada are already leveraging Illumina NGS to generate retrospective analyses of tuberculosis outbreaks[2]. In the UK, the Welcome Trust Sanger Institute used MiSeq to classify methicillin-resistant *Staphylococcus aureus* (MRSA) isolates involved in a neonatal intensive care unit at a hospital[3].

Retrospective Analysis of *Listeria* Outbreak Samples

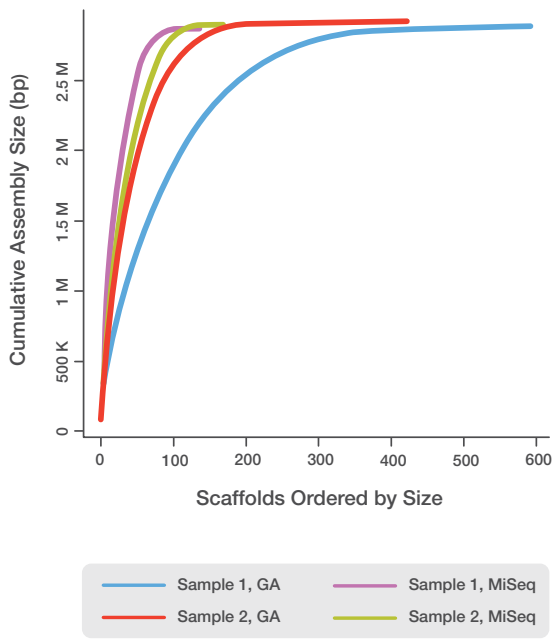
The CDC has experience with NGS, specifically with the Illumina Genome Analyzer™ system in its Biotechnology Core Facility. The rapidly falling cost and fast turnaround time of NGS sparked PulseNet's interest in learning more about the MiSeq system's capabilities for bacterial subtyping. To evaluate the system, PulseNet provided Illumina with eight *Listeria* samples from the 2011 outbreak (four from infected people and four from tainted cantaloupe). Illumina sequenced them on a MiSeq system to demonstrate its workflow, test for concordance between MiSeq and the Genome Analyzer, and determine how closely related the eight samples were to a reference strain (*L. monocytogenes* 10403S obtained from The Broad Institute[4]).

An integral part of the MiSeq system's streamlined sequencing workflow is Nextera library preparation (Figure 1). With the Nextera® Sample Prep Kit v2, sequencer-ready libraries were prepared with less than 20 minutes hands-on time. To generate the maximum amount of data for in-depth analysis, the eight samples were pooled for sequencing on a single 2 × 150 bp run at 50× coverage (~3.0 Mb genome size). Sample prep and sequencing took less than two days, followed by a day of data analysis consisting of *de novo* assembly to compare the sequencing results with those obtained with the CDC Core Facility's Genome Analyzer, resequencing data that were analyzed against the reference strain, and creation of a cluster dendrogram to illustrate sample relatedness.

Table 2: N50 Data for Listeria Samples

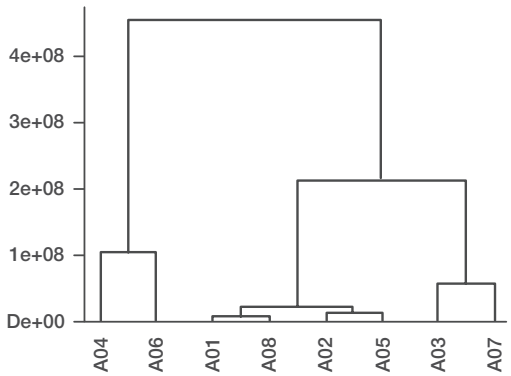
Sample	N50 (kb)	N90 (kb)
Sample #1 (Genome Analyzer IIx)	13.92 (61)	2.68 (207)
Sample #1 (MiSeq)	55.00 (15)	15.77 (54)
Sample #2 (Genome Analyzer IIx)	30.72 (28)	6.78 (106)
Sample #2 (MiSeq)	36.96 (24)	11.57 (76)

Figure 2: N50 Cumulative Scaffold Size Plot



The results from the resequencing of the outbreak samples with the MiSeq and Genome Analyzer matched. Lines with higher slope represent assemblies which are more contiguous. The MiSeq assemblies are more contiguous and therefore have higher slopes.

Figure 3: Cluster Dendrogram



Clustering dendogram of the *Listeria* samples based on SNP calling results. The clustering closely matches the results obtained with PFGE genotyping.

Two of the samples sequenced on the MiSeq were identical to two samples that the CDC sequenced on the Genome Analyzer. Only the N50 data of those four samples was compared.

MiSeq *de novo* sequencing results were concordant with the data generated by the Genome Analyzer (Table 1). The Genome Analyzer data were from a 1 × 70 bp run and the MiSeq data from a 2 × 150 run. To perform a meaningful comparison, both data sets were downsampled to the same coverage. Several assemblies were performed over a range of k-mer sizes, before choosing those with the highest N50 (50% of the assembled genome lies in contigs of this size or larger) (Table 2). MiSeq and its paired-end data provided a larger N50 scaffold size, demonstrating the value of a paired-end run for *de novo* sequencing. All assemblies had a high agreement with the reference genome.

Figure 2 shows the cumulative scaffold sizes for two MiSeq samples and two corresponding Genome Analyzer data sets. Lines with higher slope represent assemblies which are more contiguous. Both MiSeq assemblies are more contiguous and therefore have higher slopes.

Resequencing of the outbreak samples yielded > 92% alignment with the reference genome. Variant analysis filtered for Q (SNP) / Q (indel) > 20 demonstrated samples 1, 2, 5, and 8 were closely related, with two outlier groups formed by samples 4 and 6, and 3 and 7. The digital NGS data enabled the creation of a dendrogram to illustrate the hierarchical clustering of the samples (Figure 3). The results matched what PulseNet obtained viewing the analog results (bands) generated by PFGE, and were supported by the raw SNP and indel calls, and the analysis of missed/extra segments obtained in *de novo* assembly.

Summary

NGS has the potential to become a valuable, cost-effective tool for PulseNet's laboratory network, supporting its mission of public health surveillance, outbreak detection, identification, source attribution, and containment. This retrospective analysis demonstrates the potential of the MiSeq platform to generate *de novo* and resequencing data in an outbreak scenario, generating results that are comparable with the CDC Core Facility's Genome Analyzer. The MiSeq system's fast turnaround matches the speed of PFGE, while its higher resolution and accuracy enhance the ability to identify closely related bacterial isolates. Its one-touch operation supports its use in regional laboratories where ease of use is essential. Data from the MiSeq and other Illumina sequencers can also be analyzed in a cloud computing environment, such as Illumina BaseSpace™, where it can be shared in real-time with regional, national, and international public health agencies during an outbreak.

References

1. http://www.cdc.gov/listeria/outbreaks/cantaloupes-jensen-farms/index.html
2. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, et al. (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 364: 730–739.
3. http://www.sanger.ac.uk/about/press/2012/120613.html
4. http://www.broadinstitute.org/annotation/genome/listeria_group/GenomeDescriptions.html#LM_10403S_V6

Illumina, Inc. • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

© 2012 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, NuPCR, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 770-2012-021 Current as of 26 July 2012

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA