FDA U.S. FOOD & DRUG ADMINISTRATION
OFFICE OF REGULATORY AFFAIRS

Two are Better Than One: Hybrid Genome Assembly using MiSeq and MinION

Jason Neal-McKinney

[1]US Food and Drug Administration, Office of Regulatory Science, Pacific Northwest Laboratory, Applied Technology Branch. 22201 23rd Dr. SE, Bothell, WA 98021. Jason.Neal-Mckinney@fda.hhs.gov

Introduction

Whole genome sequencing (WGS) techniques allow FDA scientists to identify foodborne pathogens, determine relationships between strains, and predict the virulence and antimicrobial resistance traits of bacterial isolates. The ultimate goal of any WGS effort is to have a complete and accurate genomic sequence, where the entire genetic code of the organism is known. However, due to the limitations of each sequencing technology, most genome assemblies are fragmented into many contigs or contain assembly errors that can affect gene characterization. In this study, we improved the quality and completeness of our genome assemblies by combining data from two complementary sequencing platforms, the Illumina MiSeq and the Oxford Nanopore MinION. Raw sequence data from each instrument was uploaded to the FDA GalaxyTrakr environment, which hosts a wide variety of bioinformatics programs. QUAST and Nanostat were used to determine basic read metrics and quality, while SPAdes and Canu were used to assemble MiSeq and MinION data, respectively. Hybrid genome assembly was performed using Unicycler to combine both data types. We found that while the MiSeq data is highly accurate and useful for single nucleotide polymorphism analysis, it resulted in incomplete genome assemblies that did not include all genes. In contrast, the MinION data resulted in complete genomes that contained many errors affecting gene characterization. By combining both datatypes, we were able to generate a highly accurate and complete genome assembly that allows for analysis of every gene in the organism.

Methods

Genomic DNA was extracted from 7 *Listeria monocytogenes* isolates using the Qiagen DNEASY kit. For MiSeq, libraries were prepared using the Illlumina DNA prep kit and sequenced using paired-end 2 x 250bp reads. For MinION, libraries were prepared using the Oxford Nanopore Technologies Rapid Barcocding kit and sequenced on a single 106D Flow cell. Phylogenetic relatedness was obtained from the NCBI pathogen Detection Portal. Basic read metrics were determined using FastQC and NanoStat for Miseq and MinION data, respectively. *De novo* assemblies were generated using SPAdes for MiSeq data and Unicycler to combine both MiSeq and MinION data. Assembly metrics were determined using QUAST and the assembly graphs were visualized using Bandage. Whole genome alignment was performed using the progressiveMauve algorithm. All of the programs used were hosted in the Galaxytrakr.gov virtual environment.

Figure 1. SNP Tree of 7 *L. monocytogenes* isolates

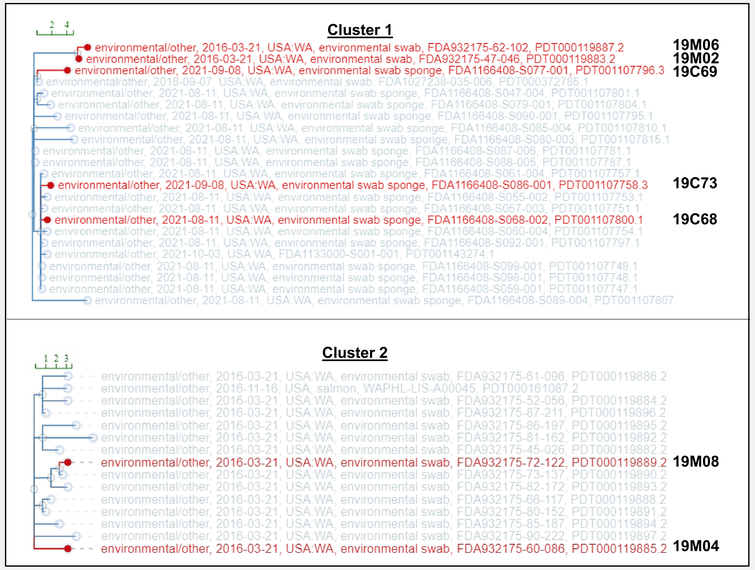


Table 1. MiSeq and MinION Read Metrics

Strain	Sequencer	Mean Read Length	Mean Read Quality	Number of Reads
19C68	MiSeq	236	36	727299
19C69	MiSeq	236.5	36.3	1813006
19C73	MiSeq	198.9	36.2	2396644
19M02	MiSeq	185.2	35.9	816386
19M04	MiSeq	173.7	34.9	1118504
19M06	MiSeq	174.4	35.2	1155587
19M08	MiSeq	181.2	36	1199700
19C68	MinION	9314.5	11.9	4138
19C69	MinION	9704.2	11.8	48784
19C73	MinION	7858.4	11.7	133865
19M02	MinION	8973.9	11.7	119543
19M04	MinION	10386	11.7	131541
19M06	MinION	10650.9	11.9	10038
19M08	MinION	9066.1	11.7	126983

Table 2. *De novo* Assembly Metrics

Strain	Sequencer	Assembler	Contigs	Total Length	Largest Contig	Circular Contigs
19C68	MiSeq	SPAdes	29	3,228,591	823,173	0
19C69	MiSeq	SPAdes	22	3,147,021	823,173	0
19C73	MiSeq	SPAdes	23	3,066,348	823,173	0
19M02	MiSeq	SPAdes	22	3,117,063	791,224	0
19M04	MiSeq	SPAdes	19	2,940,821	693,838	0
19M06	MiSeq	SPAdes	31	3,117,184	418,169	0
19M08	MiSeq	SPAdes	24	3,021,530	585,188	0
19C68	MiSeq/MinION	Unicycler	3	3,267,493	3,101,927	3
19C69	MiSeq/MinION	Unicycler	2	3,184,668	3,101,936	2
19C73	MiSeq/MinION	Unicycler	1	3,101,947	3,101,947	1
19M02	MiSeq/MinION	Unicycler	2	3,152,721	3,069,989	2
19M04	MiSeq/MinION	Unicycler	1	2,968,820	2,968,820	1
19M06	MiSeq/MinION	Unicycler	3	3,154,255	3,069,829	2
19M08	MiSeq/MinION	Unicycler	1	3,054,986	3,054,986	1

Figure 2. Assembly Graph Images

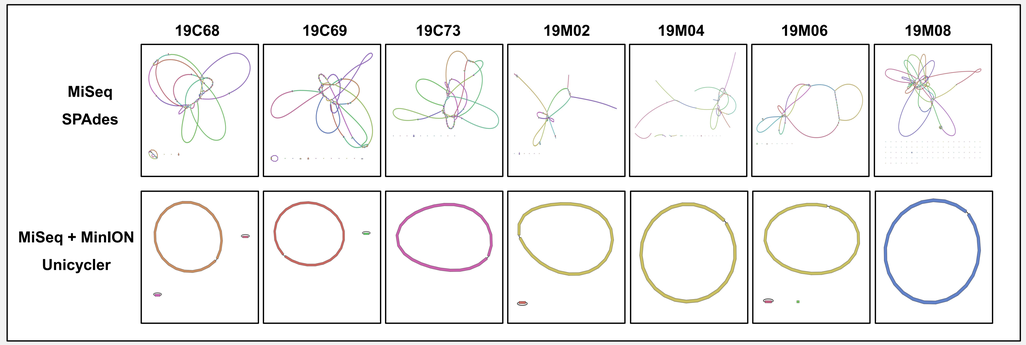
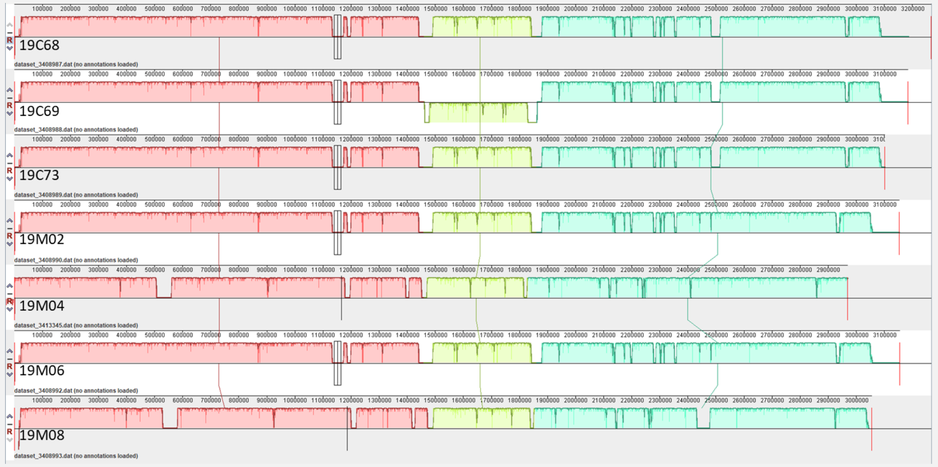


Figure 3. Whole Genome Alignments



Conclusions

- Single Nucleotide Polymorphism (SNP) analysis of the MiSeq data revealed the seven *Listeria monocytogenes* strains belong to two distinct clusters: Cluster 1 (19C68, 19C69, 19C73, 19M02, and 19M06) and Cluster 2 (19M04 and 19M08).
- *De novo* assembly of short-read data from the MiSeq using SPAdes results in fragmented genome assemblies, while the combination of MiSeq and long-read MinION data using Unicycler results in more complete genome assemblies.
- The assembly graphs of MiSeq data do not clearly show what contigs are extrachromosomal, while the chromosome and plasmids in the hybrid assemblies are circularized and separate.
- The completed genome sequences from hybrid assembly using Unicycler enable direct alignments of the whole genome that show more detail between closely related strains than the SNP analysis.
- Whole genome alignment reveals that isolates from Cluster 1 have distinct insertions and deletions relative to Cluster 2.
- Isolate 19C69 has a genomic inversion relative to other isolates in its cluster that is not apparent from SNP typing.
- The combination of long-read MinION data with the existing MiSeq data available in GenomeTrakr can be used to generate complete genome sequences useful for distinguishing closely related isolates.

Acknowledgements

All authors are FDA full-time employees. We would like to thank the FDA Pacific Northwest Laboratory for scientific assistance and support as well as the Office of Regulatory Science for review. Special thanks to the GalaxyTrakr team for hosting the bioinformatics programs used for analysis, as well as Federico Grau and Brad Tenge for local computing support.

Disclaimer

The views presented here do not necessarily reflect those of the U. S. Food and Drug Administration, nor do the authors specifically endorse the listed reagents and instruments.