Optimizing source tracking of *Listeria monocytogenes* with quasimetagenomics and integrated long and short read sequencing

Seth Commichaux (✉ Seth.Commichaux@fda.hhs.gov)

CFSAN: Center for Food Safety and Applied Nutrition <https://orcid.org/0000-0002-2162-4205>

Kiran Javkar

University of Maryland at College Park

Padmini Ramachandran

Center for Food Safety and Applied Nutrition

Niranjan Nagarajan

Genome Institute of Singapore

Denis Bertrand

Genome Institute of Singapore

Yi Chen

Center for Food Safety and Applied Nutrition

Elizabeth Reed

Center for Food Safety and Applied Nutrition

Narjol Gonzalez-Escalona

Center for Food Safety and Applied Nutrition

Errol Strain

Center for Food Safety and Applied Nutrition

Hugh Rand

Center for Food Safety and Applied Nutrition

Mihai Pop

University of Maryland at College Park

Andrea Ottesen

Center for Food Safety and Applied Nutrition

Research article

Keywords: *Listeria monocytogenes*, source tracking, pathogens

Posted Date: December 2nd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-119006/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: Rapid pathogen detection is essential for an effective public health response. The Illumina MiSeq short read sequencer has been the workhorse of many whole genome sequencing source tracking programs. However, short reads cannot span many bacterial genomic repeats, resulting in highly fragmented assemblies. Long read sequencing can span the repeats resulting in the complete assembly of bacterial genomes. With the advancing throughput and accuracy of inexpensive long read sequencing data, it is important to evaluate this resource for potential inclusion in rapid response pathogen identification protocols.

Results: Here we sought to maximize the utility of long read (Gridlon, Oxford Nanopore) and short read (MiSeq, Illumina) sequencing data for the identification of *Listeria monocytogenes* from naturally contaminated ice cream. Aliquots from the temporal enrichment (quasimetagenomes) of *L. monocytogenes* were sequenced and assembled with 10 different assembly approaches. Long read assembly tools generated genome-length contigs and a circularized 71 kbp putative *L. monocytogenes* plasmid; however, the high sequencing error rate prevented strain typing at even 150X depth of coverage. Short read assemblies provided accurate core gene strain typing after 28 hours of enrichment but were too fragmented for typing the full gene set and reconstructing a circularized genome and plasmid. Hybrid approaches demonstrated the most promising results but were biased by assembly strategy. Short read assemblies scaffolded with long reads were able to accurately strain type *L. monocytogenes* after just 24 hours of enrichment, but were still too fragmented for sequence typing the full gene set. Long read assemblies polished with short reads recovered genome-length contigs, the full complement of genes, and a circularized plasmid after 24 hours enrichment; however they could not consistently achieve the core gene strain typing accuracy of short read or short read hybrid assembly approaches.

Conclusion: These analyses demonstrate that strategic integration and optimization of microbiological (quasimetagenomic), molecular (integrated long and short reads), and bioinformatic approaches (diverse assembly strategies) can greatly improve our ability to quickly and accurately identify pathogens.

Box 1

Box 1: Assembly approaches evaluated

*For simplicity, we define a tool as a hybrid assembly approach when it uses both short and long reads whether it be short read assemblies that are scaffolded with long reads or long read assemblies that are polished with short reads.

Tool	Application	Abbreviation
MegaHit	short read metagenome assembler	short read
Redbean	long read genome assembler	long read
Canu	long read genome assembler	long read
metaFlye	long read metagenome assembler	long read
Racon	polishing long read assemblies with long reads	long read
HybridSpades	hybrid genome assembler; short read assembly first, followed by long read scaffolding	short read hybrid
Opera-MS	hybrid metagenome assembler; short read assembly first, followed by long read scaffolding either 1) de novo or 2) using reference genomes	short read hybrid
ntEdit	polishing long read assemblies with short reads	long read hybrid
Pilon	polishing long read assemblies with short reads	long read hybrid

Introduction

Rapid response, whole-genome sequencing (WGS) networks such as GenomeTrakr [1] and PulseNet [2], have revolutionized the strain typing and source attribution of bacterial, eukaryotic, and viral pathogens relevant to human and animal health. These programs have relied primarily on high throughput short-read sequencing data generated through the Illumina MiSeq platform. Accurate strain typing

of pathogens can be accomplished using short reads to quantify the number of single nucleotide polymorphisms (SNP) between an isolate and a reference genome [3]. It can also be accomplished using multilocus strain typing (MLST) [4][5]. Both methods can differentiate between very closely related strains of *Salmonella enterica*, *Listeria monocytogenes*, *Escherichia coli*, and other important human and animal pathogens. SNP-based source tracking pipelines often require the selection of an appropriate reference with respect to which SNP variants are identified and quantified. Core genome (cgMLST) and whole genome (wgMLST) MLST source tracking methods require a set of reference genomes to identify the core or complete gene set, respectively, of a species and to build databases of gene variants.

Ideally, complete genomes would be directly assembled de novo from outbreak samples for strain typing analyses; however, accurate assembly remains a challenge for short read data. Although short reads can be sequenced with an error rate of less than 1%, they cannot span many genomic repeat regions, resulting in fragmented assemblies that preclude the recovery of complete bacterial genomes [6]. In contrast, long read sequencing technologies have much higher sequencing error rates but can produce sequences that span genomic repeats, supporting the assembly of complete bacterial genomes and plasmids. The Oxford Nanopore platform has steadily advanced a financially accessible long read sequencing platform that can identify species of interest in real-time (often hours or even minutes). This technology has demonstrated utility in a range of applications, from the sequencing of human chromosomes [7], to the sequencing of Ebola virus in the field [8], and holds great promise for integration into rapid response public health initiatives. Furthermore, studies have shown that the hybrid assembly of Illumina short and Nanopore long reads can greatly improve the quality of the assemblies versus those using long reads alone, while maintaining high contiguity. For example, Nanopore long read assemblies, polished with MiSeq short reads, reconstructed the entire genomes of Shiga-toxin producing *Escherichia coli* strains [9]. Despite being of high enough quality to differentiate the strains, however, the polished genome assemblies had less accurate recovery of the cgMLST than contigs assembled using MiSeq Illumina short reads only. Another study similarly showed, for several strains of *Salmonella enterica*, that short read assembly followed by long read scaffolding more accurately reconstructed the genomes than either long read or short read assembly alone [10].

Irrespective of sequencing technology, a fundamental challenge for the source tracking of pathogens using sequencing data is the extraction of sufficient quantities of pathogen DNA. This is because pathogens frequently occur at low abundance in complex microbial communities, sometimes amongst large numbers of host cells, and in chemically challenging matrices. Current methods address this challenge by selective culture enrichment and pure colony isolation of the pathogens prior to sequencing and analysis. This approach, however, is labor-intensive and can take days to weeks to provide sufficient DNA for sequencing. Further drawbacks of this method include: (1) The method must be developed and validated for different matrices for each pathogen—a long and laborious process. (2) The recovery efficiency varies significantly depending on the background microbiota and matrix chemistry—limiting the ability to recover relevant strains. (3) Closely related non-pathogenic strains can preferentially enrich—obscuring recovery of clinically relevant strains. (4) The community dynamics during enrichment are poorly understood and non-target species often co-enrich and compete with the pathogen of interest.

Metagenomics is the direct sequencing of microbial communities [11] and, in theory, could replace culture enrichment for pathogen source tracking; however, achieving sufficient depth of coverage to assemble pathogen genomes directly from unenriched communities (metagenomes) is often prohibitively expensive. A middle ground between the direct sequencing of samples and the sequencing of isolates from selective enrichments is the sequencing of abbreviated recovery enrichments [12]. This approach is referred to as quasimetagenomics [11] and combines the strengths of both targeted culturing and direct sequencing. Short read sequencing, although widely used in metagenomics due to low error rates and high throughput, cannot assemble many of the genomic and intergenomic repeats present in environmental DNA. In contrast, long read sequencing technologies have seen far less use for metagenomics (and have not been used for quasimetagenomics before this study) due to higher sequencing error rates, lower throughput (impacting the limit of detection for specific targets), and technical biases. However, long reads can completely assemble a small subset of the bacterial genomes in metagenomic data [6, 13][14].

Here we provide the first study we are aware of, that benchmarks the utility of integrated short and long read sequencing technologies for applied quasimetagenomic pathogen source tracking. Other studies have shown that integrating short read and long read information can reconstruct complete and accurate pathogen genome assemblies from pure culture isolates [9, 10]. We sought to extend this work by testing whether we could reconstruct complete and accurate *L. monocytogenes* genomes from low species diversity quasimetagenomic samples.

Materials And Methods

Experimental design

Using long and short read sequencing technologies, we compared the performance of various assembly approaches for sequence typing *L. monocytogenes* from naturally contaminated ice cream samples that had been selectively enriched for *L. monocytogenes*. The isolation of a pure colony of *L. monocytogenes* typically requires up to 6 days of selective culture enrichment. During the selective enrichment, aliquots were collected at 4-hour intervals from 24 to 40 hours. MiSeq short read and Gridlon long read sequencing were performed on DNA from these incremental enrichments. For each time point, we assembled the sequence data using short reads alone, long reads alone, and both short and long reads. The quality of the assemblies was evaluated by comparing them to a complete *L. monocytogenes* reference strain—sequenced and assembled from PacBio data—obtained from the full enrichment protocol.

Enrichment

Ice cream samples, associated with the 2015 Blue Bell multistate listeriosis outbreak, were homogenized and added to Buffered *Listeria* Enrichment Broth (BLEB) with pyruvate according to the specifications outlined in Chap. 10 of the FDA BAM [15]. After four hours, three filter sterilized selective agents (M52) were added to achieve final concentrations of 10 mg/L acriflavin, 40 mg/L cycloheximide, and 50 mg/L sodium nalidixic acid in the BLEB. Four replicates of negative (no ice-cream) and positive controls (*L. monocytogenes* cells) were also evaluated for bacterial growth every four hours over the 40-hour enrichment.

DNA extraction and sequencing for short reads

For each of the enrichment time points (24, 28, 32, 36, and 40 hours), DNA was extracted using DNeasy Blood and Tissue kit (Qiagen) following the protocol for Gram-positive bacteria with minor modifications: 1.5 ml of the culture was pelleted (5000 × g, 15 min) and the pellet resuspended in 200 mL of enzymatic lysis buffer containing 20 mM Tris-HCl (pH-8.0), 2 mM Sodium EDTA, 1.2% Triton X-100, 20 mg/ml of lysozyme. Samples were incubated for 60 min at 37 °C. Short read libraries were prepared with Nextera Flex (Illumina) library prep kit according to the manufacturer's specifications. Libraries from enrichment time points 24, 28, 32, 36, and 40 were multiplexed along with 20 other libraries from different time points from the same study on to Illumina Miseq 2 × 250 cartridge (Illumina, CA) following manufacturer recommended protocol.

DNA extraction and sequencing for long reads

For each enrichment time point (24 to 40 hours in 4 hour intervals), 2 ml aliquots of enrichment were removed and pelletized using a benchtop Centrifuge (Eppendorf 5418 R, NY, USA) at 4000 rpm for 10 min. The pellet was resuspended in 300 ul of TE Buffer. 300 ul of the resuspended cells were loaded on the Maxwell® RSC Instrument (automated DNA extraction instrument, Madison, WI, USA) cartridge for DNA extraction. Genomic DNA was extracted using Maxwell® RSC Cultured Cells DNA Kit (Cat no: AS1260, Madison, WI, USA) on Maxwell RSC instrument following the manufacturer recommended protocol for Gram-positive bacteria.

Sequencing libraries were prepared using the ligation sequencing kit (Cat no: SQK-LSK109, Oxford Nanopore, Oxford, UK), according to the manufacturer's specifications along with Native Barcoding Expansion 1–12 (Cat no: EXP-NBD104, Oxford Nanopore, Oxford, UK) for multiplexing the samples. The libraries were multiplexed into 2 pools (Pool1: 24hr, 28hr, 32hr: Pool 2: 36hr, 40hr). The libraries were sequenced using Gridlon with Flow cell (Cat no: FLO-MIN106, Oxford, UK) following the manufacturer's recommended protocol.

L. monocytogenes reference genome

The complete *L. monocytogenes* genome (Genbank accession NZ_CP016213.1) was used as a reference for our analyses. It had previously been PacBio sequenced and assembled from an ice cream sample. The complete genome is 3,030,827 bp long with 2,984 protein-coding genes (2,710,041 bp in total length) predicted by Prokka [16] and has a GC-content of 38%. Using the cgMLST scheme developed for *L. monocytogenes* in-house at the FDA [4][5] the 1,013 cgMLST genes (1,075,554 bp) were identified and extracted using BLAST [17]. We also estimated with red [18] (only counting regions of at least 500 base pairs) that ~ 11.7% of the genome is low complexity (i.e. highly repetitive) with the longest region being 6,067 bp long.

Partitioning raw reads for comparison by sequenced base pair

By default, the Gridlon outputs the raw signal data in batches of 4,000 sequenced reads in fast5 format files [19]. Each fast5 file was converted into fastq formatted DNA sequences using Guppy for basecalling [20]. The fast base calling mode was used, which has a speed of ~ 4.6 Mbp/second. The time for the Gridlon to output a batch of reads generally ranges from 30 to 60 minutes (internal to lab) and is affected by factors such as the length and quality of the DNA fragments being sequenced. We used the first 30 batches of reads (i.e. the first 120,000 reads sequenced) and cumulatively combined them (for example, batch 1 is the first 4000 reads, batch 2 is the first 8000 reads, batch 3 is the first 12000 reads, and so on), at each enrichment time point for our analysis. So we could more directly compare assembly results across sequencing technologies, we partitioned the MiSeq raw read files into 30 corresponding batches (reads were used sequentially in the order present in the fastq files) matched to the Gridlon batches by the number of base pairs (within 500 bp which is the approximate length of forward and reverse reads). The total number of sequenced bases ranged from a minimum of 231 million at 24 hours of enrichment to a maximum of 533 million at 36 hours of enrichment.

Raw read statistics and reference genome coverage

Raw read statistics were collected for the 30 batches of reads per enrichment time point, including: mean per base quality score, number of reads, number of base pairs, read length distribution, and estimated sequencing error rate. To estimate the sequencing error rate we first mapped the short reads and long reads to the *L. monocytogenes* reference genome with Bowtie2 [21] and MiniMap2 [22], respectively, using default settings. We then counted and normalized the number of mismatches, insertions, and deletions for the mapped reads with respect to the reference genome. For the Gridlon reads, we provide an estimated range for the sequencing error rate because MiniMap2 is a local, as opposed to a global, read alignment tool. The range is based on whether soft-clipping of the read alignments is included as sequencing error (maximum estimate of error) or not (minimum estimate of error). Insertions, deletions, and mismatches were only counted for the aligned portion of the reads i.e. excluding the soft-clipped regions. The read mappings were used to estimate the breadth and depth of coverage (DOC) of the *L. monocytogenes* reference genome.

Assembling the sequenced reads

Short reads and long-reads from each of the 30 batches were assembled per enrichment time point. The short reads were assembled using MegaHit [23] with default settings and scaffolded with MetaCarvel [24]. The long reads were assembled using Canu [25], Redbean [26], and metaFlye [27] with default settings. The Redbean assemblies were polished with MiniMap2 and SAMtools [28] following the tutorial for Redbean on its GitHub page. Unlike metaFlye, which is a long read metagenome assembler, Canu and Redbean are not designed for metagenomic assembly. However, we chose these assemblers for these comparative analyses as they are frequently used long-read genome assemblers. All of the metaFlye assemblies were polished, using the long reads, with Racon [29]. HybridSpades [30] and Opera-MS [31] (with and without reference genome scaffolding) were used for short read hybrid assembly – short read assembly first followed by scaffolding with the long reads. Opera-MS was chosen because it is a metagenome assembler, while HybridSpades was chosen because it is a popular genome assembler. Pilon [32] and ntEdit [33] were used for long read hybrid assemblies – long read assembly first with metaFlye followed by short read polishing. Each tool was run with 12 CPUs.

Assembly statistics

Assembly statistics per batch at each enrichment time point were gathered using Quast [34] including: the runtime; the total assembly length; the number of contigs; the longest contig; the N50 of the metagenomic assemblies; the NG50 of the assembled *L. monocytogenes* contigs with respect to the reference genome; and, the number of insertion/deletions/mismatches in the assembled *L. monocytogenes* contigs with respect to the reference genome. We also calculated the BLAST distance between the set of detected cgMLST genes and the reference cgMLST (total length 1,075,554 bp) as well as the cgMLST synteny relative to the reference genome. Additionally, the BLAST distance between the complete set of genes (i.e. wgMLST) in the reference (total length 2,710,041 bp) and those detected in the assemblies was also measured.

Here, we define the BLAST distance as the number of mismatches, insertions, and deletions in the BLAST alignment between the reference genes and the assembled genes. Preferably, we would have calculated the edit distance between the reference genes and the genes found in the assemblies, but correct identification of complete genes, especially in noisy long read assemblies, is complicated; therefore, our results only approximate the true edit distance between the reference and the assemblies. The BLAST distance will either be equal to or greater than the edit distance.

Taxonomic classification

The contigs from the short read assembly MegaHit and long read assembly metaFlye assemblies were taxonomically classified with Kraken [35] and the MiniKraken database using default settings. A species was considered present if a minimum of 5,000 nt of contigs were annotated as that species.

The proportion of reads mapped to the *L. monocytogenes* genome was used as a proxy for the relative abundance of *L. monocytogenes* in the samples.

It is known from other work evaluating the genomes connected to this outbreak that there were at least three strains of *L. monocytogenes* [36]. To identify whether the samples used for our analysis had multiple strains of *L. monocytogenes*, we mapped the short reads to the reference genome with Bowtie2 and performed variant detection analysis with Pilon. The number of loci marked ambiguous by Pilon was recorded as potential evidence of multiple strains.

Results

For our analysis, we sought to benchmark the performance of short read, long read, and hybrid assembly tools for quasimetagenomic pathogen source tracking. We evaluated assembly tools that were developed for metagenomic assemblies: MegaHit for short read assembly, metaFlye for long read assembly, and Opera-MS for hybrid assembly. We also evaluated popular tools developed for long read genome assembly (Canu and Redbean) and a hybrid genome assembler, HybridSpades. Additionally, we evaluated 3 polishing tools: Pilon, ntEdit (both were used to polish long read assemblies with short reads), and Racon (was used to polish long read assemblies with long reads). For the purpose of benchmarking, the assembly approaches were broadly categorized into 4 groups: 1) short read assembly (MegaHit), 2) long read assembly (metaFlye, Canu, Redbean, metaFlye + Racon), 3) short read hybrid assembly i.e. short read assembly first followed by scaffolding with long reads (HybridSpades, Opera-MS), 4) long read hybrid assembly i.e. long read assembly first followed by polishing with short reads (metaFlye + Racon + ntEdit, metaFlye + Racon + Pilon).

The resulting assemblies were compared using metrics including: total assembly size, N50, NG50, longest contig assembled, and the number of assembly errors. Additionally, the strain-typing utility of the approaches (measured by cgMLST and wgMLST recovery) was evaluated using a high-quality, complete (PacBio sequenced) *L. monocytogenes* reference genome isolated from the same enrichment.

Quasimetagenomic samples were generated from ice cream that was naturally contaminated with *L. monocytogenes* [36]. At selected intervals, aliquots of the culture were collected and prepared for short and long read sequencing. Sequence data were assembled using several assembly tools (previously described) representative of short read, long read, and hybrid assembly approaches. Assemblies were evaluated for their completeness and their utility for high-resolution phylogenetic source tracking. With these results, we could contrast the performance of the various tools at incremental timepoints in the culture process and examine the culture enrichment community dynamics. The resulting data and assembly metrics allowed us to provide specific recommendations regarding innovations for bacterial pathogen source tracking.

Characteristics of the sequencing data

The mean read length across enrichment time points ranged from 174–198 nucleotides for Illumina MiSeq and 1,923–4,445 nucleotides for Oxford Nanopore Gridlon. The longest Gridlon reads ranged from 48,588–69,402 nucleotides (Table 1). For the Gridlon, there was a general increase in the mean and maximum read length as the enrichment time increased. Furthermore, the reads that mapped to the *L. monocytogenes* reference genome had a longer mean and maximum length compared to the rest of the reads across all enrichment time points (Supplementary Fig. 1). The putative *L. monocytogenes* reads also had a much lower mean GC content (38%) compared to the rest of the reads, which varied slightly from 49–54% across enrichment time points (Supplementary Fig. 2).

Table 1
Nanopore read length and sequencing error statistics

Enrichment time (hours)	Mean read length	Max. read length	Average quality score	Min. est. sequencing error rate	Max. est. sequencing error rate
24	1923	48588	21.8	7%	18%
28	2721	55258	22.9	6%	17%
32	3268	57233	22.8	7%	16%
36	4445	62426	23.2	6%	13%
40	4129	69402	23.2	6%	13%

The sequencing error rate was 0.03% for the MiSeq reads and between 6.3–18% for the Gridlon reads. For the Gridlon, we provided a range for the sequencing error rate which was estimated by including or excluding the soft-clipping of the reads (ends of the reads not included in the alignment range). Each read is thus assigned two error estimates: an upper estimate of error that treats the unaligned portion of the read as an error, and a lower estimate that relies solely on the errors identified within the aligned range. Insertions, deletions, and mismatches were only counted for the aligned portion of the reads i.e. excluding the soft-clipped regions. For the long reads, 29.6%, 25.4%, and 45% of the errors were due to mismatches, insertions, and deletions, respectively. For the MiSeq, the sequencing error rate and effective read lengths were relatively uniform across samples. For the Gridlon, the sequencing error rate decreased from 24 hours (18–7%) to 40 hours (13–6.3%) which correlated with a general increase in the mean per base quality score over time from 21.83 to 23.19 from 24 to 40 hours of enrichment, respectively.

General metagenome assembly statistics

Most assembly approaches had a mean runtime of approximately 40 minutes or less (Table 3) per sample. Canu was the only exception with a mean runtime of 98 minutes per sample. The fastest assembly approach was Redbean with a mean runtime of just one minute (Supplementary Fig. 3).

Table 3
Mean assembly statistics (across all batches and time points) for each of the assembly approaches

Assembly tool	Runtime	Total assembly length	Number of contigs	N50	Longest contig	Misassemblies	Mismatches per 100kbp	Indels per 100kbp
metaFlye (long read)	40.6	4291417	27	3056133	3056133	0.4	1.0	481.3
Canu (long read)	98	3470967	21	1754979	2071553	10.8	26.6	265.1
Redbean (long read)	1	3474503	35	2123769	2131343	2.6	31.8	291.3
MegaHit (short read)	32.8	7972605	7315	97577	672182	1.2	3.8	1.2
metaFlye_Racon (long read)	41.6	4261624	27	3039238	3039238	1	3.4	73.7
HybridSpades (short read hybrid)	22.6	11681048	19285	112850	686270	0	7.0	2.2
OperaMS_no_ref (short read hybrid)	12.2	10340211	13921	105382	655220	2.2	1.5	2.3
OperaMS_ref_cluster (short read hybrid)	13.6	10363273	13913	205943	1919416	4.2	1.8	2.3
metaFlye_Racon_Pilon (long read hybrid)	41.6	4271759	27	3041086	3041086	1	1.9	3.3
metaFlye_Racon_ntEdit (long read hybrid)	41.6	4274358	27	3041440	3041440	1	1.4	4.0

We characterized the contiguity of the resulting assemblies by measuring the total assembly length (Supplementary Fig. 4), number of contigs (Supplementary Fig. 5), N50, (Supplementary Fig. 6), and longest contig assembled (Supplementary Fig. 7). The mean values, across all samples, for each contiguity metric are described in Table 3. Approaches that begin with assembling short reads (short read and short read hybrid assemblies) contrasted substantially with those that begin with assembling long reads (long read and long read hybrid assemblies) having consistently longer total assembly lengths, orders of magnitude more contigs, lower N50s, and shorter longest contigs. With the enrichment of *L. monocytogenes* (and the corresponding loss in the number of species from the original community), however, there was a general decrease in the number of contigs and assembly size.

As expected, the long read and long read hybrid assemblies consistently had the highest N50 values and the longest contigs—often near the reference genome length for *L. monocytogenes* (~ 3 Mbp). In contrast, the short read and short read hybrid assemblies had low N50 values and longest contigs that were consistently shorter (often by orders of magnitude), and showed little to no improvement beyond 60X depth of coverage. Opera-MS, using reference-guided scaffolding, was the main exception, assembling contigs of 2 Mbp or more at all enrichment time points. Amongst the long read assembly tools, the metagenome assembler metaFlye consistently produced the highest N50 values with the longest contigs nearest to the length of the *L. monocytogenes* reference genome; however, the differences between long read assembly tools decreased with enrichment.

Taxonomic composition of the quasimetagenomic samples

Taxonomic annotation of the assembled sequences predicted a total of 10 species based on all of the samples, platforms, and assembly methods, with the number of species predicted per sample ranging from 2 to 10 (Fig. 1). The number of species predicted decreased with enrichment time, with short read and short read hybrid assemblies predicting more species than long read and long read hybrid assemblies. *L. monocytogenes* and *Rothia mucilaginosa* were the only species detected at all time points, across all assemblies, with *L. monocytogenes* consistently being the most abundant species. The most frequently detected and closely related species to *L. monocytogenes* was *Bacillus cereus* (both organisms are from the order *Bacillales*).

The abundance of *L. monocytogenes* increased with enrichment time; however, the estimates of abundance from the MiSeq and Gridlon differed slightly (Table 2). At 24 hours of enrichment, the proportion of reads from *L. monocytogenes* was 33% for MiSeq and ~ 60% for Gridlon. After 40 hours of enrichment, ~ 92% of the MiSeq reads mapped to the *L. monocytogenes* reference genome compared to ~ 97% of Gridlon reads.

Table 2 Percent of Illumina and Nanopore reads that map to the <i>L. monocytogenes</i> reference genome		
Enrichment time (hour)	Illumina (reads mapped with Bowtie2)	Nanopore (reads mapped with MiniMap2)
24	33	60
28	68	88
32	75	94
36	88	97
40	92	97

We could not rule out the presence of multiple *L. monocytogenes* strains, but if multiple strains were present they varied by very few nucleotides. Whole genome variant analysis revealed 71 variant loci at 24 hours enrichment and 14 variant loci at 40 hours enrichment. As the sequencing error rate also decreased over time, it is uncertain whether these represent true variants or just errors.

Quality of the assembled *L. monocytogenes* genome within the metagenome

The most contiguous recovery of the *L. monocytogenes* genome, as measured by the mean NG50 across time enrichment points, was by long read and long read hybrid assembly approaches (Fig. 2). Amongst the long read assemblers, metaFlye most consistently assembled genome-length contigs for *L. monocytogenes* at the earliest time points and at the lowest depth of coverage compared to Canu or Redbean. For example, the mean NG50's for the assembled *L. monocytogenes* contigs, across all batches of sequenced reads and enrichment time points, for Canu, Redbean and metaFlye were 1,535,966 bp, 1,568,760 bp and 2,490,733 bp, respectively. For this

reason only the metaFlye assemblies were used for the long read hybrid assemblies. The long read hybrid approaches (Racon, Pilon, ntEdit) slightly decreased the mean NG50 of the metaFlye assemblies, 2,477,272 bp, 2,478,715 bp, 2,478,772 bp, respectively. The short read Megahit assemblies had the smallest mean NG50 at 162,346 bp. The de novo short read hybrid assemblies only slightly increased the mean NG50 (HybridSpades (431,211 bp) and Opera-MS without reference-guided scaffolding (375,881 bp)). However, the Opera-MS assemblies that used reference-guided scaffolding had an order of magnitude increase in the mean NG50 over Megahit i.e. 1,414,301 bp.

Only the long read assemblers were able to assemble genome-length contigs (over 3 million bp) for *L. monocytogenes*. The earliest time point at which a genome-length contig, corresponding to *L. monocytogenes*, was reconstructed was at 24 hours of enrichment with metaFlye (batch 18, 7.2×10^4 sequenced nanopore reads, 33X depth of coverage of the *L. monocytogenes* genome) and Canu (batch 22, 8.8×10^4 sequenced nanopore reads, 40X depth of coverage of the *L. monocytogenes* genome), and 28 hours with Redbean (batch 16, 6.4×10^4 sequence nanopore reads, 47X depth of coverage of the *L. monocytogenes* genome). The genome length contigs, irrespective of the long read assembly tool used, were almost always longer than the actual reference genome, mostly due to over-circularization of the assembly by a read length or less, sometimes by tens of thousands of base pairs. Additionally, each long read assembler recovered a circularized 71 kbp putative *L. monocytogenes* plasmid that was always found broken into multiple contigs in the short read assemblies. The assembled plasmid had BLAST best hits (using NCBI's nt database) to other known *L. monocytogenes* plasmids but did not host any known resistance or virulence genes.

The assembled contigs annotated as *L. monocytogenes* were assessed for their quality with respect to the reference genome using Quast (Fig. 3). The number of misassemblies and mismatches per 100 kbp varied more by tool than sequencing strategy; however, there was a pronounced difference between assembly approaches for indels per 100 kbp. The mean number of misassemblies ranged from 10.8 (Canu) to 0 (HybridSpades). The mean number of mismatches per 100 kbp per enrichment time point ranged from 31.8 (Redbean) to 1.2 (metaFlye). In terms of indels, long read assembly approaches had a high mean number of indels per 100 kbp, ranging from 265 (Canu) to 481 (metaFlye). The combination of metaFlye with Racon substantially reduced the number of indels to 74 per 100 kbp. Combining short read and long read information with long read hybrid assembly approaches further reduced the number of indels to ~ 3 per 100 kbp. Short read assembly/short read hybrid assembly approaches had the lowest indel rate of around 1 to 2 per 100 kbp.

Strain-typing: Identification of cgMLST and wgMLST

Accurate identification of the core genes and complete set of genes is critical for accurate cgMLST and wgMLST strain typing, respectively. Multilocus sequence typing (MLST) uses the combination of alleles found for target genes to sequence type strains of a given species. At all enrichment time points, for both short and long reads, there was 100% breadth of coverage of the *L. monocytogenes* reference genome by the reads, and the depth of coverage ranged from ~ 1X to ~ 150X. We assessed the strain typing accuracy of the assembly tools by calculating the BLAST distance for the cgMLST (Fig. 4) and wgMLST (Fig. 5) between the reference genome and the assemblies. We define the BLAST distance as the number of mismatches, insertions, and deletions in the BLAST alignment between the reference genes and the assembled genes. In general, we observed that the cgMLST was most accurately identified in the short read hybrid and short read assemblies, while the wgMLST was most accurately identified in the long read hybrid assemblies. The short read hybrid and short read assemblies were too fragmented for accurate typing of the wgMLST. Conversely, the long read hybrid assemblies were noisier than the short read hybrid and short read assemblies, typing the cgMLST with slightly less accuracy and precision.

For the cgMLST (Fig. 4), we never observed a BLAST distance smaller than 5 for any assembly approach. This means that the reference genome, which was sequenced and assembled at the end of the complete enrichment process, had 5 nucleotide differences in the core genes compared to the *L. monocytogenes* assemblies from the quasimetagenomes. These differences imply that the *L. monocytogenes* reference genome represents a closely related strain or that there were assembly errors.

The short read hybrid assembly approaches recovered the cgMLST with BLAST distance 5 at the earliest time point: HybridSpades at 24 hours enrichment and batch 22 (8.8×10^4 sequenced nanopore reads), corresponding to 40X (long reads) and 19X (short reads) depth of coverage of the *L. monocytogenes* reference genome; Opera-MS, both with and without reference-guided scaffolding, at 24 hours enrichment and batch 28 (1.12×10^5 sequenced nanopore reads), corresponding to 50X (long reads) and 25X (short reads) depth of coverage of the *L. monocytogenes* reference genome. Megahit assemblies attained a BLAST distance of 5 after 28 hours of

enrichment and batch 11 (corresponds to 4.4×10^4 sequenced nanopore reads), corresponding to 28X depth of coverage of the *L. monocytogenes* reference genome.

Overall, for the cgMLST, the long read assembly approaches performed poorly, never achieving a BLAST distance of less than 2,000 even with 158X depth of coverage of *L. monocytogenes*. Polishing metaFlye assemblies with Racon substantially improved the long read assemblies, achieving a minimum BLAST distance of 609. Long read hybrid assembly with Pilon achieved a BLAST distance of 5 at 28 hours of enrichment and batch 14 (5.6×10^4 sequenced nanopore reads), which corresponded to 36X (short reads) and 38X (long reads) depth of coverage of the *L. monocytogenes* reference genome; however, it did so less consistently than short read or short read hybrid approaches (Fig. 4). Long read hybrid assembly with ntEdit performed similarly to but not as well as that with Pilon.

For the wgMLST (Fig. 5), the long read hybrid approaches were the most accurate, with Pilon outperforming ntEdit. Pilon achieved a BLAST distance of 132, the best observed for any tool, at 28 hours enrichment and batch 14 (5.6×10^4 sequenced nanopore reads), corresponding to 36X (short reads) and 38X (long reads) depth of coverage of the *L. monocytogenes* reference genome. The mean BLAST distance across enrichment time points was 699 for Pilon and 798 for ntEdit. None of the other assembly approaches attained this level of accuracy; for reference, the next best tool, metaFlye + Racon, had a mean BLAST distance of 2,991. The long read hybrid approaches also showed high consistency as batches of sequenced reads were added with a median difference from batch-to-batch of BLAST distance 137 for Pilon and 140 for ntEdit.

Batch effects: The effect of adding batches of sequenced reads on assembly quality and strain typing

In addition to accuracy, the precision with which a strain can be typed is of great importance for pathogen detection. We observed that the addition of sequence data (batches of sequenced reads) resulted in a wide range of batch-to-batch variability between assembly tools for the cgMLST and wgMLST strain typing of *L. monocytogenes* (Fig. 6).

For cgMLST typing, the most consistently performing tools from batch-to-batch were Opera-MS (without reference guided scaffolding), HybridSpades, MegaHit, and metaFlye + Racon + Pilon; the median difference in BLAST distance between added batches of sequenced reads for these tools was 0, 1, 1, and 5, respectively, across all enrichment time points. All assembly approaches had a median difference in BLAST distance of less than 50 except Opera-MS with reference guided scaffolding (121), Canu (1,183) and Redbean (10,930).

For wgMLST typing, the level of variability as batches of sequenced reads were added was an order of magnitude greater than that for cgMLST typing. The most consistent tools for wgMLST typing were HybridSpades, metaFlye + Racon + Pilon, metaFlye + Racon + ntEdit, metaFlye + Racon, and metaFlye—the median difference in BLAST distance between batches was 132, 137, 140, 183, and 207, respectively, for these tools. All assembly approaches had a median difference in BLAST distance of less than 1,000, with the exceptions of Opera-MS with reference guided scaffolding (1,019), Canu (4,758), and Redbean (32,655).

Besides batch-to-batch variability, we also observed that increased depth of coverage did not always correlate with improved performance in assembly metrics. Some metrics plateaued at a certain depth of coverage. For example, short read assemblies did not see much improvement in the longest contig assembled beyond 30X coverage (at 28 hours enrichment and 30X depth of coverage, the longest contig was 695,760 nt, while at 40 hours enrichment and 151X depth of coverage, the longest contig was 695,778 nt). In other cases, the performance of assembly approaches actually decreased with increased depth of coverage. For example, the wgMLST BLAST distance for the metaFlye + Racon + Pilon assemblies increased from 132 after 28 hours enrichment to 153 after 40 hours enrichment, despite an increase of 100X depth of coverage of the *L. monocytogenes* genome for both short and long reads.

Discussion

This study sought to benchmark the performance of short read, long read, and hybrid assembly approaches for strain-typing *L. monocytogenes* in quasimetagenomic samples. Quasimetagenomic sequencing is the middle ground between traditional culture-based and culture-independent metagenomic approaches, requiring far less culture enrichment time than what is currently needed to obtain pure-isolate colonies for whole genome sequencing.

Our study showed that short read sequencing detects more species than long read sequencing for the same number of sequenced base pairs, perhaps because there are fewer reads being sequenced from the microbial community. The result was a higher percentage of the long reads coming from the most abundant organisms in the sample. Furthermore, there was evidence that read

length varies by species for the long reads as the *L. monocytogenes* mean read length was much higher than that for the reads from the rest of the species in the sample, at all enrichment time points.

Summary of tool performance by assembly approach

The short read, MegaHit assemblies accurately recovered the complete cgMLST given 28 hours of enrichment, but could not sequence type the wgMLST, nor fully reconstruct the *L. monocytogenes* genome or plasmid.

The long read assemblies could recover the complete genome, on a single contig with the correct cgMLST synteny, and a circularized putative plasmid of *L. monocytogenes*; however, the high level of noise from sequencing errors prevented accurate strain-typing of *L. monocytogenes* even with 158X depth of coverage. Amongst the long read assembly tools, metaFlye (the only metagenomic assembler) assembled the complete genome and the putative plasmid of *L. monocytogenes* the most consistently and accurately, with the fewest assembly errors and the least depth of coverage. This highlights that the best assembly strategies differ for genomic and metagenomic data. Furthermore, polishing the metaFlye assemblies with Racon greatly increased the assembly quality demonstrating that the currently available long read assemblers are not fully utilizing the information contained in the long reads.

Amongst the short read hybrid assembly approaches, we observed different advantages with different tools. HybridSpades and Opera-MS (without reference-guided scaffolding) performed similarly. Both were able to accurately and precisely type the cgMLST with BLAST distance 5 (the lowest value observed in our experiments) as early as 24 hours enrichment—although HybridSpades achieved BLAST distance 5 the earliest, i.e. with the least depth of coverage. They also produced assemblies that were less fragmented than the short read assemblies with over double the mean NG50 values for *L. monocytogenes*. Opera-MS when run with the reference-guided scaffolding option, consistently had much higher N50 and NG50 values (sometimes assembling contigs near the length of the *L. monocytogenes* reference genome) and increased the typing accuracy of the wgMLST compared to the other two approaches. However, the assemblies also had more assembly errors and typed the cgMLST with less precision and accuracy. This highlights how using reference genomes to scaffold assemblies can increase contiguity, but can also introduce assembly errors if there are structural differences between the reference and the genome being assembled. Irrespective of the short read hybrid assembly approach, the assemblies were too fragmented to recover a complete genome or plasmid and the wgMLST typing accuracy was not competitive with the long read hybrid assembly approaches.

The two long read hybrid assembly approaches (metaFlye + Racon + Pilon and metaFlye + Racon + ntEdit) performed similarly to metaFlye in terms of the general assembly statistics (N50, total assembly length, longest contig assembled, number of contigs), but with far fewer insertion/deletions and mismatches in the assemblies. Amongst all the assembly approaches tested, metaFlye + Racon + Pilon most accurately and precisely typed the wgMLST. The long read hybrid assemblies were not able to type *L. monocytogenes* via the cgMLST as precisely nor accurately as the short read and short read hybrid assemblies.

The trade-off in performance between short read hybrid and long read hybrid assemblers reflected the difference in the initial assembly strategy. For short read assembly first approaches, the indel error rate was much lower, resulting in more accurate cgMLST typing, but the assemblies were also much more fragmented. For long read assembly first approaches, the indel rate was higher, preventing accurate cgMLST typing; however the assemblies had much higher contiguity, resulting in the complete reconstruction of the *L. monocytogenes* genome and plasmid with the most accurate wgMLST typing. These differences, despite both assembly strategies using short and long read data, highlights the opportunity for the development of hybrid metagenomic assembly approaches which incorporate both their strengths. One such approach might assemble the long reads and short reads separately first and then map the short read assembly contigs to the long read contigs to attain a more contiguous and accurate assemblies.

Lastly, we often observed great variability in the performance of tools as batches of sequenced reads were added and assembled. This highlights the differential sensitivity of assembly approaches to additional sequence data and suggests that assembly tools for strain typing should be developed and benchmarked for consistency.

Despite the many advances in long read sequencing over the past decade, comprehensive benchmarking studies are still needed, especially for metagenomic (or quasimetagenomic) data, to assess the strengths and weaknesses of the sequencing technologies and supportive software. Our data provides a valuable benchmark from real samples with few species and a known pathogen with a sequenced reference genome. A limitation of our study is the lack of closely related species and strains, in the samples. The closest related species to *L. monocytogenes* that we detected in our samples was *Bacillus cereus*, which is only related at the order level. The

presence of closely related species and/or strains would have allowed for the assessment of higher resolution challenges for assembly such as inter-/intragenomic repeats. We expect that such features would have resulted in less contiguous, noisier assemblies with more misassemblies such as chimeric contigs.

Long read sequencing technologies, like the Nanopore GridION, have demonstrated utility for assembling complete bacterial genomes –although improvements in throughput and a decrease in the sequencing error rate are needed for high quality performance in metagenomic contexts. In the context of quasimetagenomics, throughput is an important consideration as there is an inverse relationship that needs to be minimized between the amount of sequencing necessary to attain sufficient depth of coverage and the amount of enrichment time.

Rapid, efficient assembly of complete genomes and plasmids will underpin important advances in our understanding of genome biology and evolution. Of extreme importance will be the description of the still poorly characterized plasmids, viruses, and mobile elements that play complex roles in pathogenicity and the spread of antimicrobial resistance. The additional information provided by integrated methodologies from microbiology to bioinformatics will also provide unprecedented speed and resolution for strain-typing.

Conclusion

Here we benchmarked the performance of short read, long read, and hybrid assembly approaches for source attribution (via high-resolution MLST) of a pathogen from a food matrix (ice-cream naturally contaminated with *L. monocytogenes*). Our analyses demonstrate that the combination of quasimetagenomics and hybrid assembly can accurately strain type and reconstruct a complete *L. monocytogenes* genome after only 24 hours of enrichment (current approaches to pure isolate culturing of *L. monocytogenes* and genome sequencing can take 6 days). We observed that hybrid assembly approaches provided the most precise and accurate strain typing. Hybrid approaches provided assemblies with much greater contiguity than short read assemblies and far fewer errors than long read assemblies. Despite their advantages, there were strategy specific weaknesses and strengths for the hybrid approaches. Short read hybrid assembly approaches achieved the most rapid and accurate cgMLST typing, but were not able to assemble complete genomes. Long read hybrid assembly approaches were able to assemble the complete genome of *L. monocytogenes* and most accurately detected the wgMLST, but could not match the accuracy of short read hybrid assembly approaches for cgMLST typing. The discrepancy between hybrid assembly strategies highlights areas for future innovation.

The ability to source track pathogens as quickly as possible during outbreaks is a high priority for public health labs. Pure culture isolation for the targeted sequencing of pathogen genomes is the current gold standard, but is slow, sometimes taking weeks. Quasimetagenomics is a new approach to pathogen source attributions with demonstrated success [12]. As the long read sequencing error rate and cost decrease and as the throughput continues to increase, we can expect the utility of long read sequencing technology for source tracking to grow. Rapid and accurate assembly of whole genomes from quasimetagenomic data will greatly increase the speed and resolution with which source-tracking can occur. Additionally, it will provide vital data to describe plasmid and resistome geography. Our microbiological, molecular, and bioinformatic protocols must also continue to evolve and optimize technological advances. We are not far from widespread public health integration of real-time pathogen detection and source tracking directly from quasimetagenomic data.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The short and long reads are available on NCBI in Bioproject PRJNA630588. The *L. monocytogenes* reference genome is available on Genbank under accession NZ_CP016213.1. The code and data used for our analysis is available on GitHub at https://github.com/SethCommichaux/Long_read_short_read_comparison.

Competing interests

The authors declare that they have no competing interests.

Funding

Funding was supported by the Office of Regulatory Science, Center for Food Safety and Applied Nutrition, and the FDA. Kiran Javkar was supported by the University of Maryland, Joint Institute for Food Safety and Applied Nutrition (JIFSAN) through cooperative agreement #5U01-FD001418, provided by the U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition (FDA, CFSAN).

Authors' Contributions

Commichaux and Javkar performed bioinformatic analyses, wrote the manuscript and created figures. Ottesen and Ramachandran designed the experiment. Ottesen, Ramachandran, Chen and Gonzalez-Escalona performed microbiological and molecular lab work. Reed created figures and organized data. Nagarajan and Bertrand ran Opera-MS assemblies and analyses. Commichaux, Javkar, Ottesen, Ramachandran, Chen, Gonzalez-Escalona, Strain, Rand and Pop edited the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

We would like to acknowledge the qMGS pathogen detection efforts of the MetagenomeTrakr initiative of FDA. Thanks to our collaborators from the National Antimicrobial Resistance Monitoring System (NARMS), the Office of Regulatory Affairs (ORA), and at the Center for Veterinary Medicine (CVM).

References

1. Allard MW, Strain E, Melka D, Bunning K, Musser SM, Brown EW, et al. Practical Value of Food Pathogen Traceability through Building a Whole-Genome Sequencing Network and Database. *J Clin Microbiol*. 2016;54:1975–83.
2. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, the CDC PulseNet Task Force. PulseNet: The Molecular Subtyping Network for Foodborne Bacterial Disease Surveillance, United States. *Emerging Infectious Diseases*. 2001;7:382–9. doi:10.3201/eid0703.017303.
3. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, et al. CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Computer Science*. 2015;1:e20. doi:10.7717/peerj-cs.20.
4. Pightling AW, Petronella N, Pagotto F. The *Listeria monocytogenes* Core-Genome Sequence Typer (LmCGST): a bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiol*. 2015;15:224.
5. Pettengill JB, Pightling AW, Baugher JD, Rand H, Strain E. Real-Time Pathogen Detection in the Era of Whole-Genome Sequencing and Big Data: Comparison of k-mer and Site-Based Methods for Inferring the Genetic Distances among Tens of Thousands of *Salmonella* Samples. *PLoS One*. 2016;11:e0166162.
6. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol*. 2020. doi:10.1038/s41587-020-0422-6.
7. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*. 2018;36:338–45.
8. Hoenen T, Groseth A, Rosenke K, Fischer RJ, Hoenen A, Judson SD, et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg Infect Dis*. 2016;22:331–4.
9. González-Escalona N, Allard MA, Brown EW, Sharma S, Hoffmann M. Nanopore sequencing for fast determination of plasmids, phages, virulence markers, and antimicrobial resistance genes in Shiga toxin-producing *Escherichia coli*. *PLoS One*. 2019;14:e0220494.
10. Chen Z, Kuang D, Xu X, González-Escalona N, Erickson DL, Brown E, et al. Genomic analyses of multidrug-resistant *Salmonella* Indiana, Typhimurium, and Enteritidis isolates using MinION and MiSeq sequencing technologies. *PLoS One*. 2020;15:e0235641.
11. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*. 2004;68:669–85.

12. Ottesen A, Ramachandran P, Chen Y, Brown E, Reed E, Strain E. Quasimetagenomic source tracking of *Listeria monocytogenes* from naturally contaminated ice cream. *BMC Infect Dis.* 2020;20:83.
13. Hamner S, Brown BL, Hasan NA, Franklin MJ, Doyle J, Eggers MJ, et al. Metagenomic Profiling of Microbial Pathogens in the Little Bighorn River, Montana. *Int J Environ Res Public Health.* 2019;16. doi:10.3390/ijerph16071097.
14. Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. *Sci Rep.* 2020;10:13588.
15. United States. Food and Drug Administration. Bacteriological Analytical Manual. Aoac International; 1995.
16. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30:2068–9. doi:10.1093/bioinformatics/btu153.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
18. Girgis HZ. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC Bioinformatics.* 2015;16:227.
19. [No title]. <https://store.nanoporetech.com/media/it-requirements-configure-gridion-capex.pdf>. Accessed 29 Oct 2020.
20. Basecalling using Guppy. https://timkahlke.github.io/LongRead_tutorials/BS_G.html. Accessed 30 Oct 2020.
21. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature Methods.* 2012;9:357–9. doi:10.1038/nmeth.1923.
22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–100.
23. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 2015;31:1674–6.
24. Ghurye J, Treangen T, Fedarko M, Hervey WJ 4th, Pop M. MetaCarvel: linking assembly graph motifs to biological variants. *Genome Biol.* 2019;20:174.
25. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. *Genome Research.* 2017;27:722–36. doi:10.1101/gr.215087.116.
26. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17:155–8.
27. Kolmogorov M, Rayko M, Yuan J, Pevzner P. metaFlye: scalable long-read metagenome assembly using repeat graphs. doi:10.1101/637637.
28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
29. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research.* 2017;27:737–46. doi:10.1101/gr.214270.116.
30. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics.* 2016;32:1009–15.
31. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol.* 2019;37:937–44.
32. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9:e112963.
33. Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, Isabel N, et al. ntEdit: scalable genome sequence polishing. *Bioinformatics.* 2019;35:4430–2.
34. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* 2013;29:1072–5.
35. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology.* 2014;15:R46. doi:10.1186/gb-2014-15-3-r46.
36. Ottesen A, Ramachandran P, Reed E, White JR, Hasan N, Subramanian P, et al. Enrichment dynamics of *Listeria monocytogenes* and the associated microbiome from naturally contaminated ice cream linked to a listeriosis outbreak. *BMC Microbiol.* 2016;16:275.

Figures

Figure 1

Taxonomic classification of all the quasimetagenomic data from each of the enrichment time points. For the sake of clarity, only short read MegaHit and long read metaFlye assemblies were plotted (short read assembly results mirrored short read hybrid assemblies and long read assemblies mirrored long read hybrid assemblies). A) The total bp of contigs per species (must have a minimum of 5,000 bp) classified by Kraken. B) Species in sample, excluding *L. monocytogenes*, *R. mucilaginosa* and unclassified sequences - highlights how the short read assemblies capture more species than the long read assemblies.

Figure 1

Taxonomic classification of all the quasimetagenomic data from each of the enrichment time points. For the sake of clarity, only short read MegaHit and long read metaFlye assemblies were plotted (short read assembly results mirrored short read hybrid assemblies and long read assemblies mirrored long read hybrid assemblies). A) The total bp of contigs per species (must have a minimum of 5,000 bp) classified by Kraken. B) Species in sample, excluding *L. monocytogenes*, *R. mucilaginosa* and unclassified sequences - highlights how the short read assemblies capture more species than the long read assemblies.

Figure 2

The NG50 for the assembled *L. monocytogenes* contigs at each of the enrichment time points for each assembly approach. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 2

The NG50 for the assembled *L. monocytogenes* contigs at each of the enrichment time points for each assembly approach. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 3

The quality of assembled contigs annotated as *L. monocytogenes* with respect to the reference genome using Quast. The number of mismatches, insertion/deletion (indels), and misassemblies per 100 kbp at each of the enrichment time points for each assembly approach. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 3

The quality of assembled contigs annotated as *L. monocytogenes* with respect to the reference genome using Quast. The number of mismatches, insertion/deletion (indels), and misassemblies per 100 kbp at each of the enrichment time points for each assembly approach. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 4

BLAST distance between the reference genome cgMLST and assembly as a function of total sequenced base pairs. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 4

BLAST distance between the reference genome cgMLST and assembly as a function of total sequenced base pairs. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 5

BLAST distance between the reference genome wgMLST and assembly as a function of total sequenced base pairs. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 5

BLAST distance between the reference genome wgMLST and assembly as a function of total sequenced base pairs. (Abbreviations: SR = short read, LR = long read, HY = hybrid)

Figure 6

Consistency of assembly approaches. Median batch-to-batch difference in A) cgMLST and B) wgMLST BLAST distances across enrichment time points.

Figure 6

Consistency of assembly approaches. Median batch-to-batch difference in A) cgMLST and B) wgMLST BLAST distances across enrichment time points.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure1.pdf](#)
- [SupplementaryFigure1.pdf](#)
- [SupplementaryFigure2.pdf](#)
- [SupplementaryFigure2.pdf](#)
- [SupplementaryFigure3.pdf](#)
- [SupplementaryFigure3.pdf](#)
- [SupplementaryFigure4.pdf](#)
- [SupplementaryFigure4.pdf](#)
- [SupplementaryFigure5.pdf](#)
- [SupplementaryFigure5.pdf](#)
- [SupplementaryFigure6.pdf](#)
- [SupplementaryFigure6.pdf](#)
- [SupplementaryFigure7.pdf](#)
- [SupplementaryFigure7.pdf](#)