microorganisms

MDPI

Article

Complete Whole Genome Sequences of *Escherichia coli* Surrogate Strains and Comparison of Sequence Methods with Application to the Food Industry

Dustin A. Therrien [1], Kranti Konganti [2], Jason J. Gill [1], Brian W. Davis [3], Andrew E. Hillhouse [2], Jordyn Michalik [1], H. Russell Cross [1], Gary C. Smith [1], Thomas M. Taylor [1,]* and Penny K. Riggs [1,]*

[1] Department of Animal Science, Texas A&M University, College Station, TX 77843-2471, USA; therrienda@gmail.com (D.A.T.); jason.gill@tamu.edu (J.J.G.); jordyn.michalik@tamu.edu (J.M.); hrcross@tamu.edu (H.R.C.); gary.smith@tamu.edu (G.C.S.)

[2] Texas A&M Institute for Genome Sciences and Society, MS 2470, College Station, TX 77843-2470, USA; k.kranti.neu@gmail.com (K.K.); hillhouse@tamu.edu (A.E.H.)

[3] Department of Veterinary Integrated Biosciences, Texas A&M University, College Station, TX 77843-4461, USA; bdavis@cvm.tamu.edu

* Correspondence: matt_taylor@tamu.edu (T.M.T.); riggs@tamu.edu (P.K.R.); Tel.: +979-862-7015 (P.K.R.)

Abstract: In 2013, the U.S. Department of Agriculture Food Safety and Inspection Service (USDA-FSIS) began transitioning to whole genome sequencing (WGS) for foodborne disease outbreak- and recall-associated isolate identification of select bacterial species. While WGS offers greater precision, certain hurdles must be overcome before widespread application within the food industry is plausible. Challenges include diversity of sequencing platform outputs and lack of standardized bioinformatics workflows for data analyses. We sequenced DNA from USDA-FSIS approved, non-pathogenic *E. coli* surrogates and a derivative group of rifampicin-resistant mutants (rif^R) via both Oxford Nanopore MinION and Illumina MiSeq platforms to generate and annotate complete genomes. Genome sequences from each clone were assembled separately so long-read, short-read, and combined sequence assemblies could be directly compared. The combined sequence data approach provides more accurate completed genomes. The genomes from these isolates were verified to lack functional key *E. coli* elements commonly associated with pathogenesis. Genetic alterations known to confer rif^R were also identified. As the food industry adopts WGS within its food safety programs, these data provide completed genomes for commonly used surrogate strains, with a direct comparison of sequence platforms and assembly strategies relevant to research/testing workflows applicable for both processors and regulators.

Keywords: bacterial surrogate; *Escherichia coli*; whole genome sequence; short reads; long reads; closed genome; high throughput sequencing

check for updates

Citation: Therrien, D.A.; Konganti, K.; Gill, J.J.; Davis, B.W.; Hillhouse, A.E.; Michalik, J.; Cross, H.R.; Smith, G.C.; Taylor, T.M.; Riggs, P.K. Complete Whole Genome Sequences of *Escherichia coli* Surrogate Strains and Comparison of Sequence Methods with Application to the Food Industry. *Microorganisms* **2021**, *9*, 608. https://doi.org/10.3390/microorganisms9030608

Academic Editor: Christina L. Swaggerty

Received: 12 February 2021
Accepted: 10 March 2021
Published: 16 March 2021

1. Introduction

Over recent decades, the landscape of food safety has undergone paradigm shifts as technological advancements in genomics enabled implementation of numerous measures for ensuring a safe and secure food supply [1]. Various methodologies (e.g., pulsed-field gel electrophoresis (PFGE), serotyping, phage typing, multi-locus sequence typing (MLST), etc.) have been utilized for identification and characterization of foodborne pathogens at the clinical level. While these techniques played invaluable roles in food safety, technological limitations prevent the resolution necessary for differential identification of closely related bacterial strains. Moreover, illnesses attributed to foodborne pathogens continue to persist and are estimated to be responsible for approximately 48 million illnesses (1 in 6 people), 128,000 hospitalizations, and 3000 deaths each year within the United States [2–6]. However, rapid technological advancements and drastic reductions in cost have made applications such as whole genome sequencing (WGS) via high throughput sequencing (next generation sequencing (NGS) and 3rd generation sequencing) appealing alternatives [7–9].

Currently, WGS is achieved via two types of sequencing methods that can be distinguished by the length of sequence fragments or "read lengths" (i.e., short- and long-read) produced [8,10]. Short-read sequencing platforms, such as those manufactured by Illumina, utilize massively parallel sequencing that yields read lengths of about 100 to 300 base pairs (bps) with a high level of accuracy. Typically, error rates in nucleotide identification (base calling) are less than 1% and result in 95% coverage of most bacterial genomes [8,10–13]. The short-read approach allows a researcher to make comprehensive estimations regarding the total number of genes present within the organism of interest, their classification in relation to other species, and the overall relatedness of their distinct gene sets to other organisms. While highly informative and effective for comparative gene-based studies, this technique is inadequate for producing sequences that span long repetitive genomic regions and large areas that are prone to rearrangement (e.g., deletions, insertions, repeats, and inversions). This limitation frequently results in incomplete genomic assemblies (draft genomes) of contiguous segments (contigs) that are oriented incorrectly or contain other structural errors due to repetitive elements and other genome features [8,10,12–14]. In contrast, long-read sequencing platforms can generate read lengths ranging from ~1000 bps to hundreds of kilobases in a single read. Unfortunately, the increased sequence length is offset by a significant reduction in sequence accuracy, with base calling error rates ranging from ~5–40% [8,11,13–17].

Despite the differences among WGS technologies, sequencing-based approaches consistently provide greater resolution and discriminatory power for distinguishing closely related bacterial species compared to previous methods, thus improving foodborne pathogen surveillance systems and trace back investigations [4,5,18–22]. WGS datasets can be simultaneously used in multiple investigative analyses (e.g., subtyping, antibiotic resistance profiling, virulence genetic markers, screening of mobile genetic markers, etc.) or stored for future analyses. For these reasons, WGS is being adopted by federal regulatory and public health related entities (e.g., Center for Disease Control (CDC), Food and Drug Administration (FDA), USDA-FSIS) as one of the primary methods for surveillance, outbreak investigations, and the tracing of transmission routes of foodborne pathogens [8,18,22,23].

The technological benefits of WGS support adoption by government and regulatory agencies, but certain aspects must be addressed before WGS can be widely incorporated as routine screening within the food industry, if at all. Arguably, one area of greatest difficulty pertaining to this technology is the abundant diversity in workflows that exist for processing and sequencing of the samples, as well as bioinformatic analyses and interpretation of large volumes of data. Genomic technologies have undergone rapid advancements that enabled innovations and accessibility but have also resulted in a large variety of preparatory workflow procedures, sequencing platforms with diverse utility, and innumerous bioinformatic analytical tools [8,10,13,18,22,24,25].

The overall objective of this project was to produce high quality, complete genome assemblies for a group of USDA-FSIS-approved non-pathogenic *E. coli* surrogates (ATCC BAA-1427, BAA-1428, BAA-1429, BAA-1430, and BAA-1431) and annotate any virulence factor genes or subunit genes present in the genomes. To also provide direct comparison of the technologies, we conducted comparative analyses of the short-read and long-read sequences for direct comparison to the hybrid approach. We also annotated the genes conferring rifampicin resistance in derivative isolates at the same time. The USDA-FSIS has previously supported the use of non-pathogenic surrogate organisms for the validation of in-plant intervention strategies to reduce the presence of foodborne pathogens. This model of intervention is highly beneficial because it allows one to determine the efficacy of a current intervention strategy without inherent risk of contaminating testing equipment or facilities. This group of surrogates is of particular interest because they been shown to possess similar properties to pathogenic *E. coli* and *Salmonella enterica* and are widely used in contemporary research and intervention validation [26–31]. Despite their widespread use, genetic information for these strains is not readily available. Thus, the data presented here contribute to the existing body of knowledge regarding sequencing approaches for detection of genes associated with pathogenesis and antibiotic resistance. In addition, these data provide completed genomes for these widely used surrogates that will be an invaluable resource for processors and regulatory officers in differentiating these strains from pathogenic strains of *E. coli* and supporting decision-making for the incorporation and application of WGS for food safety applications.

2. Materials and Methods

2.1. Bacterial Surrogates

Five non-pathogenic *E. coli* biotype I strains (isolates BAA-1427, BAA-1428, BAA-1429, BAA-1430, and BAA-1431) were obtained from the American Type Culture Collection (ATCC; Manassas, VA, USA) and revived according to ATCC guidance [32]. The surrogates originated as isolates from cattle hides at facilities in the Department of Animal Science at Iowa State University (Ames, IA, USA) [29,33]. The strains were confirmed to lack antibiotic resistance or a subset of known virulence factors by the *E. coli* Reference Center of Penn State University (University Park, PA, USA) and underwent further toxin testing via the application of commercial kits as well as tissue culture testing (African green monkey kidney (Vero) cells) by the depositor [33]. Isolate details are described in tables available at the ATCC website [32,34]. These *E. coli* isolates were propagated twice in 5.0 mL of tryptic soy broth (TSB; Becton, Dickinson and Co., Sparks, MD, USA) (24 h, 35 °C), and then grown on tryptic soy agar (TSA; Becton, Dickinson and Co., Sparks, MD, USA) slants, TSA Petri plates, TSA + rifampicin (100.0 mg/L; TSA-R) plates, and MacConkey agar (MAC; Becton, Dickinson and Co., Sparks, MD, USA) Petri plates (24 h, 35 °C). Following overnight incubation, colonies of parent *E. coli* isolates grown on the TSA slants and streaked on plates were verified as rifampicin-sensitive or rif^{R}. API® 20E (bioMérieux, Inc. N.A., Durham, NC, USA) tests were used to identify organisms as *E. coli* according to manufacturer guidance. Additionally, including in the sequencing experiment were three rif^{R} mutants, coded BAA-1427 rif^{R}, BAA-1428 rif^{R}, and BAA-1430 rif^{R} [35,36]. Following verification of parent strain identities, these mutants were prepared from isolates BAA-1427, BAA-1428, and BAA-1430 by the T.M. Taylor lab (Texas A&M University, College Station, TX, USA) and made available for this project. The rif^{R} strains were verified as described above with respect to *E. coli* identification, and via overnight growth on TSA-R media (24 h, 35 °C).

2.2. DNA Extraction and Quantification

For DNA extraction, *E. coli* (parents, rif^{R} mutants) were propagated from working stocks incubated in 5.0 mL TSB (24 h, 35 °C) as previously described. TSA streak plates were created from each bacterial isolate and incubated likewise (24 h, 35 °C). A single bacterial colony was selected and grown in 5.0 mL TSB (24 h, 35 °C) for DNA extraction. The samples were centrifuged at 10,000 g for 2 min and cell pellets were frozen at −80 °C until used. A phenol/chloroform DNA extraction protocol was used to isolate genomic DNA from cell pellets [37]. The extraction procedure was modified with the substitution of 1-bromo-3-chloropropane (BCP; Molecular Research Center Inc., Cincinnati, OH, USA) for 24:1 chloroform:isoamyl alcohol prior to ethanol precipitation [38]. DNA samples were quantified via spectrophotometry (NanoDrop ND-1000, Thermo Fisher Scientific, Waltham, MA, USA), visualized by electrophoresis through a 1.0% agarose SFR gel (AMRESCO, Solon, OH, USA) in a 1× Tris/Borate/EDTA (TBE) buffer solution, and stained with SYBR green (Sigma-Aldrich Corp., St. Louis, MO, USA).

2.3. Genomic Sequencing

Bacterial genomic DNA was sequenced at the Texas A&M Institute for Genome Sciences and Society (TIGSS) core facility (College Station, TX, USA) by Illumina MiSeq and Oxford Nanopore MinION gene sequencing platforms (Illumina, Inc., San Diego, CA, USA and Oxford Nanopore Technologies, Oxford, UK). The derivative group of rif^R mutants were sequenced only via MiSeq. Prior to sequencing, the bacterial DNA was re-quantified via the Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) as recommended by the Illumina MiSeq and Oxford Nanopore MinION library preparation kit protocols. Libraries were prepared with the Nextra XT v2 library preparation kit (Illumina, Inc., San Diego, CA, USA) for the Illumina MiSeq platform, and the Rapid Barcoding Kit (SQK-RBK004, Oxford Nanopore Technologies, New York, NY, USA) for the MinION. Quality of all sample libraries was evaluated via the Agilent 2200 TapeStation (Agilent, Santa Clara, CA, USA) prior to sequencing

2.4. Genome Assembly

Upon completion of sequencing reactions, the raw sequence data were downloaded from the Illumina BaseSpace (Illumina, Inc., San Diego, CA, USA) and Oxford Nanopore Metrichor (Oxford Nanopore Technologies, Oxford, UK) cloud-based storage systems and uploaded onto the TIGSS High Performance Computing Cluster for further processing. The sequence data produced by the MinION were converted from the FAST5 to FASTQ format via the Oxford Nanopore Albacore v2.0.1 base caller [39]. Sequence quality was assessed via FastQC, and low-quality sequence data and the adapter sequences were removed with Trimmomatic v0.32 [40,41]. The SPAdes software tool (v3.13.0) was used to generate a short-read assembly from the MiSeq data, and the Canu v2.0 single-molecule sequence assembler was used to generate long-read assembly from the MinION data [42,43]. Once assembled the contigs for each bacterial sample were screened and those contigs that were <1000 bps (MinION), <500 bps (MiSeq), possessed low coverage scores, and/or were poorly associated with *E. coli* species were removed from the assemblies. After low-quality contigs had been removed, sequence statistics were calculated for each sample, with overall rates of coverage being calculated via the BEDTools software [44].

2.5. Polishing and Error Correction

Raw unfiltered MiSeq reads and the Canu FASTA long-read assemblies were combined into hybrid assemblies using the Unicycler genomic assembler [45]. During this process, the generated hybrid assemblies underwent various cycles of polishing and error correction using the integrated Pilon software tool v1.23 [46]. Following this the degree of completeness of each hybrid genome was assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO) software v4 and was compared with the lineage enterobacteriales (composed of 216 species and 781 orthologs) [47]. To further close these genomes, each was processed using the reference-guided contig ordering and orienting tool (RaGOO) with the *E. coli* K12 substr. MG1655 (NC_000913.3) reference genome [48]. Lastly, for the samples that were not reduced to a single contig, analysis was conducted via BLAST [49] to align the nucleotide sequences of the surplus contigs to known sequence to identify their origins (BLASTn).

2.6. Virulence Factor Screening and Verification

Serotyping and MLST for the three assemblies of each bacterial surrogate was determined with the open access SeroTypeFinder 2.0 and MLST 2.0 software [50,51]. The generated assemblies for each of the *E. coli* surrogates were analyzed by translated BLAST analysis (BLASTx) against a dataset of *E. coli* virulence factors extracted from the Virulence Factor Database (VFDB) using an e-value cutoff of 10^{-5} (Table S1) [52–55]. The genome of the known non-pathogenic *E. coli* str. Nissle 1917 (NZ CP022686) was used as a control to filter spurious hits by subtraction of BLASTx hits shared between the surrogates and Nissle with the remaining factors undergoing further investigation. Analyses were conducted on the Texas A&M University Center of Phage Technology (CPT, College Station, TX, USA) Galaxy instance [56].

The remaining detected virulence factors were examined to confirm the presence of complete genes and/or gene modules as appropriate. Individual bacterial contigs were opened with Sanger Artemis (v. 18.0.0, Wellcome Sanger Institute, Cambridge, UK) and the regions containing suspected virulence determinants based on BLASTx coordinates were manually annotated and their protein sequences compared to those of known functional virulence factors by BLASTp to determine if they were complete and free of alterations that may render them non-functional [57]. The percent identities for each of the potential pathogenic elements found within the hybrid assemblies were calculated using the Sørensen-Dice coefficient [58,59].

$$SDC = 2|x \cap\ y|/|x| + |y| \tag{1}$$

The rif^{R}-mutants were excluded from this analysis, as it was expected that their matches would correspond with their surrogate parent strains.

2.7. Detection of known rpoB Rifampicin Resistance Mutations

The *rpoB* DNA sequences of the parental surrogates BAA-1427, BAA-1428, and BAA-1430 assemblies (i.e., long-read, short-read, and hybrid), and their corresponding short-read rif^{R}-mutants BAA-1427 rif^{R}, BAA-1428 rif^{R}, and BAA-1430 rif^{R} counterparts were compared to that of *E. coli* str. K-12 substr. MG1655 (NC_000913.3) via BLASTn to detect mutations commonly associated with rifampicin resistance [60].

3. Results

3.1. Comparison of Sequence Assembly Statistics

Assembly summaries for the MinION and MiSeq assemblies were calculated and compared along with their Serotypes and MLSTs (Tables 1 and 2). Sequence generated from the Oxford Nanopore MinION platform resulted in read lengths that were approximately 10-fold longer than read outputs from the Illumina MiSeq sequencer (Tables 1 and 2). The longer read lengths enabled assembly of sequence reads into fewer and longer contiguous stretches (contigs) and resulted in greater overall genome coverage for each bacterial sample (Table 1). The draft assemblies produced from MinION data resulted in a large singular contig for each assembly (4–5 Mbs) with a subset of smaller contigs averaging 1kb in size. In contrast, the MiSeq platform resulted exhibited greater uniformity in the size distribution of contigs, with the largest ranging from ~300–500 kbs with a steady decline in size to the smallest contig which was ~70 bps (Figures S1 and S2). These results are consistent with expected ranges for each platform, as a consequence of the unique chemistry and mechanisms for each technology. However, the total assembled lengths for each bacterial genome differed by only 100-300 kb between the MinION and MiSeq sequencing platforms, reflecting a 2–6% difference among surrogate counterparts (Tables 1 and 2).

Table 1. Long-read Oxford Nanopore MinION assembly sequence statistics.

Bacterial Strains	O Type	H Type	MLST [1]	Contigs	Assembled Length	Largest Contig	Average Coverage
BAA-1427	-	4	n/a	74	5,034,864 bps	4,743,343 bps	323.673×
BAA-1428	154	16	n/a	67	5,050,340 bps	4,806,641 bps	311.819×
BAA-1429	166	12	n/a	20	4,856,504 bps	4,816,131 bps	362.642×
BAA-1430	28ac/42	21	n/a	19	5,217,837 bps	5,022,067 bps	310.567×
BAA-1431	-	4	n/a	34	4,982,422 bps	4,753,397 bps	306.167×

[1] MLST (Multi-locus Sequence Typing)–types could not be determined due to imperfect matches.

Table 2. Short-read Illumina MiSeq assembly sequence statistics.

Bacterial Strains	O Type	H Type	MLST	Contigs	Assembled Length	Largest Contig	Average Coverage
BAA-1427	-	4	10	91	4,825,300 bps	434,834 bps	51.211×
BAA-1428	154	16	165	127	4,758,825 bps	319,570 bps	57.141×
BAA-1429	166	12	10	87	4,739,915 bps	523,910 bps	60.601×
BAA-1430	28ac/42	21	278	103	5,009,161 bps	421,121 bps	49.422×
BAA-1431	-	4	10	91	4,829,685 bps	404,666 bps	47.608×

3.2. *Hybrid Assembly Statistics and Analysis*

The MinION and MiSeq assemblies were combined to improve the overall genome assembly of each surrogate. For each hybrid assembly, summary statistics were calculated and serotypes and MLSTs were identified for comparison with the long- and short-read counterparts. Total lengths of the hybrid assemblies for all five *E. coli* surrogates increased when compared with the MiSeq assemblies and slightly decreased when compared to that of the MinION assemblies (Tables 1–3). In considering the total number of contigs and the overall completeness of the genomes, significant improvements were observed in the hybrid assemblies in (Tables 1–3). In most cases the genomes were reduced to a single contig. The remaining additional contigs observed for two of the surrogates (BAA-1428 & BAA-1430) were identified via BLASTn to be residual fragments of existing plasmids. Additionally, when the hybrid genomes were compared with the lineage enterobacteriales (216 species and 781 orthologs) within BUSCO, each sample's genome was reported to be between ~99.8 and 99.9% complete (Table 3). Lastly, the hybrid assembly's quality was further improved compared with the other assemblies as it underwent multiple rounds of polishing via Pilon which resulted in numerous corrections within each genome (Table 3). GenBank Genomes database accession numbers [61] for each genomic assembly are included in Table 3, and each sample was annotated via the automated NCBI prokaryotic genome annotation pipeline [62].

Table 3. Hybrid assembly sequence statistics.

Bacterial Strains	O Type	H Type	MLST	Pilon [1]	BUSCO [2]	Contigs	Assembled Length (bps)	Largest Contig (bps)	Average Coverage	GenBank Accession No.
BAA-1427	-	4	10	6	99.9%	1	4,886,306	4,886,306	152×	CP063979
BAA-1428	154	16	165	5	99.8%	2	4,876,786	4,870,024	151×	CP063956-CP063967
BAA-1429	166	12	10	4	99.9%	1	4,812,017	4,812,017	186×	CP063969
BAA-1430	28ac/42	21	278	8	99.9%	5	5,106,612	4,988,672	138×	CP063970-CP063974
BAA-1431	-	4	10	6	99.9%	1	4,889,455	4,889,455	135×	CP063958

[1] Indicates the number of rounds of error correction each assembly underwent during Pilon processing.[2] Indicates the predicted completeness of each assembly generated by BUSCO (Benchmarking Universal Single-Copy Orthologs) after comparison to the lineage enterobacteriales.

For three of the bacterial genomes (BAA-1427, BAA-1429, and BAA-1431), each assembly was closed and reduced to a single observable contig that was within the range of a standard *E. coli* genome. The BAA-1428 genome contained one contig that was comparable in size with the three completed genomes, and a smaller 6762 bps. From BLASTn analysis, the smaller, non-chromosomal contig was found to be identical (100% coverage) to several plasmid sequences existing in public databases: *Salmonella enterica* serovar Newport plasmid pSNE1-1926 (CP025235.1) (6761 bps), *Salmonella enterica* serovar 1,4(5),12:i- plasmid p11-0813.1 (CP039594.1) (6760 bps), and *Salmonella enterica* serovar Enteritidis plasmid p4.4 (MG948564.1) (6760 bps). Of these, the proposed plasmid differed the most from *Salmonella enterica* serovar 1,4(5),12:i- plasmid p11-0813.1 (CP039594.1) by only 50 nucleotide alterations that existed primarily between nucleotides 1201–1315. Both *Salmonella enterica* serovar Newport plasmid pSNE1-1926 (CP025235.1) and *Salmonella enterica* serovar Enteritidis plasmid p4.4 (MG948564.1) possessed a nucleotide shift (A→G) at nucleotide 1371 when compared with the proposed BAA-1428 plasmid. Additionally, the proposed plasmid was compared with the other plasmids, they all possessed deletions within a region of low-complexity sequence (i.e., homopolymeric guanines) that spans between nucleotides 426–436. With the only notable differences existing within this region of low-complexity sequence and at nucleotide 1371 (A→G) it could not be determined if the proposed BAA-1428 plasmid was more similar to the *Salmonella enterica* serovar Newport plasmid pSNE1-1926 (CP025235.1) or *Salmonella enterica* serovar Enteritidis plasmid p4.4 (MG948564.1).

The result for BAA-1430 however was enigmatic when compared with the others as not only was it on average ~300 kbps larger than the other assembled genomes in total length but despite all further processing, remained at five observable contigs. Of these the largest was 4,988,672 bps in overall size which is more comparable to the other genomes. Four smaller contigs that were present ranged from 96,846, 9368, 6077, and 5649 bps in length. When BLASTn analysis was performed on these remaining non-chromosomal contigs it was found that the second contig (96,846 bps) displayed the highest genetic identity to the *E. coli fergusonii* plasmid pRHB23-C01_2 (CP057566.1; 99.74% identity, 83% coverage). The third contig (9368 bps) most resembled the *Serratia liquefaciens* plasmid pS12 (CP048786.1; 99.97% identity, 94% coverage. The fourth contig (6077 bps) shared a 100% identity and 100% coverage with the *E. coli* plasmid pRHB08-C23_3 (CP057955.1). Lastly, the fifth and smallest contig (5649 bps) revealed a 99.83% identity and 96% coverage when compared with an unnamed plasmid previously associated with *E. coli* strain RHB13-C21 (CP055721.1).

3.3. Virulence Factor Presence/Absence Determination and Characterization

The MinION, MiSeq, and hybrid genome assemblies from each of the five *E. coli* surrogates encoded genes associated with a subset of predicted regulatory protein adherence factors (Table 4). However, the genomes lacked many of the necessary genes that encode vital structure elements/subunits necessary for assembly of the full protein complexes, thus rendering these adherence factors non-functional (Table 4). The genomes assembled from MinION sequence were of lower resolution, containing multiple indels and higher errors rates that significantly reduced statistical confidence for detection of many of the adherence factor sequences that were examined, in comparison with those generated by the MiSeq and hybrid assemblies (Table 4). However, the MinION, MiSeq, and hybrid genome assemblies indicate that strains BAA-1427, and BAA-1431 encode complete cytolethal distending toxin (CDT) A, B, and C, and cytotoxic necrotizing factor 1 (CNF1) (Table 4). The percent identities of each of the identified pathogenesis factors for each hybrid assemblies were calculated with the Sørensen-Dice coefficient [58,59]. Predicted amino acid sequences identified as CDT A, -B, and -C within both surrogate sequences were 56.03%, 69.87%, and 40.56% similar to their functional CDT A, -B, and -C counterparts (GenBank: CAD48849.1, CAD48850.1, and CAD48851.1), respectively [63]. Additionally, the CNF1-like amino acid sequence in BAA-1427 and BAA-1430 possessed 53.48% percent identity to functional CNF1 (GenBank: CAA50007.1) [64].

Table 4. Virulence attributes observed in bacterial surrogates. Subunits of virulence factors that were detected in each strain are indicated. An e-value limit of <0.00001 was adopted as a cut-off for protein identity (Blastx analysis). GenBank accession numbers for the virulence factors and their corresponding subunits within this table are provided in Table S1.

Virulence Factors	BAA-1427	BAA-1428	BAA-1429	BAA-1430	BAA-1431
Bundle-forming pili subunits (BFP)	bfpB $^{*(-)}$, bfpE $^{*(-)}$, bfp$^{HI(-)}$	-	-	bfpB $^{IH(-)}$, bfpE $^{IH(-)}$, bfpH $^{IH(-)}$	bfpB $^{*(-)}$, bfpE $^{*(-)}$, bfpH $^{I(-)}$
Plasmid-encoded regulator (Per)	-	-	perC/bfpW $^{*(-)}$	perC/bfpW $^{*(-)}$	-
Cytolethal distending toxin (CDT)	cdtA $^{*(56.03\%)}$, cdtB $^{*(68.87\%)}$, cdtC $^{*(40.56\%)}$	-	-	-	cdtA $^{*(56.03\%)}$, cdtB $^{*(68.87\%)}$, cdtC $^{*(40.56\%)}$
Adhesive fimbriae	csnA $^{IH(-)}$, cswA $^{I(-)}$	csnA$^{I(-)}$, cswA$^{IH(-)}$	csnA $^{I(-)}$, cswA $^{IH(-)}$	cfaB $^{*(-)}$, cooA $^{*(-)}$, csbA $^{IH(-)}$, csnA $^{*(-)}$, cswA $^{IH(-)}$	csnA $^{IH(-)}$, cswA $^{I(-)}$
Cytotoxic necrotizing factor 1 (CNF1)	cnf1 $^{*(53.48\%)}$	-	-	-	cnf1 $^{*(53.48\%)}$
P fimbriae	-	-	papE $^{*(-)}$, papG $^{*(-)}$, papJ $^{*(-)}$	-	-

$^{(\%)}$ Indicates the percent identity (%ID) that the protein subunit shares with its virulent counterpart.$^{(-)}$ indicates that a %ID was not calculated because the virulence factor was not present or lacked required subunits. ^{I} Gene was detected in the assembly from Illumina MiSeq. ^{H} Gene was detected in the hybrid assembly. * Gene was detected in every dataset.

3.4. Detection of Known rpoB Rifampicin Resistance Mutations

The *rpoB* DNA sequences of the parental surrogates BAA-1427, BAA-1428, and BAA-1430 assemblies (i.e., long-read, short-read, and hybrid), and their corresponding short-read rif^R-mutants BAA-1427 rif^R, BAA-1428 rif^R, and BAA-1430 rif^R were compared to that of *E. coli* str. K-12 substr. MG1655 (NC_000913.3) to gauge each method's utility for enabling detection of known mutations that confer rifampicin resistance (Table S2). When screened, it was found that the three BAA-1427 assemblies (parent strains) and the BAA-1427 rif^R assembly (mutant child strain) shared a silent mutation (A206 to A), while the rif-resistant strain contained an additional L533 to P mutation. The BAA-1428 genomes and BAA-1428 rif^R shared a silent mutation (T486 to T), and BAA-1428 rif^R also possessed a mutation in S512 to P. Additionally, the parent BAA-1430 genomes and the BAA-1430 rif^R genome shared a series of silent mutations (P489 to P, L623 to L, and G846 to G) when compared to *E. coli* K-12. Lastly, in addition to those silent mutations the BAA-1430 rif^R mutant possessed an additional mutation of H526 to Y. However, it is of note that while the MinION assemblies did contain the same mutations as the MiSeq, hybrid, and rif-resistant assemblies, they also contained a large number of additional indels and were ultimately deemed unsuitable for the reliable identification of rif^R–associated single-nucleotide polymorphisms (SNPs) (Table S2). While elements known to confer rif^R could be detected in the MinION assemblies, this would not be a reliable approach for detecting novel mutations.

4. Discussion

We conducted WGS of a group of USDA-approved non-pathogenic *E. coli* surrogates via two popular NGS technologies and, also performed short-read sequencing on rif^R derivatives that exist for three of them. Our objective was to generate and characterize complete genome sequences for these important resources. At the same time, we used the opportunity to directly compare two common sequencing platforms and evaluate their usefulness for identification of potential pathogenic elements or known SNPs that confer rifampicin resistance. Both sequencing methods enabled production of draft genome assemblies for each bacterial strain, although key differences were apparent-notably in the distribution of contig size between the two platforms.

Despite producing draft genomes that typically contained less than 100 contigs of quality sufficient for comparative genomic analysis, with some exception to the MinION genomes due to high error rates, a complete, closed genome was not produced by either method alone (Tables 1 and 2). However, sequence from the MinION enabled assemblies for each sample in which a single contig comprised ~94–99% of the total assembly length (Table 1). Consistent with previous findings, when the MinION and MiSeq assemblies were utilized in producing hybrid de novo assemblies, the unique strengths of each method combined to overcome their individual limitations [12,45,65–67]. The combined hybrid MinION and MiSeq assembly resulted in drastic quality improvements in each of the bacterial genomes assemblies (Table 3). The hybrid assembly was similar in overall length but had greatly reduced contigs and improved quality for bacterial assembly. Analysis of each hybrid for completeness (via BUSCO using lineage enterobacteriales), indicated that each genome assembly was ~99.8–99.9% complete (Table 3). Overall, the hybrid assemblies proved to be superior for closing the bacterial genomes and provide an invaluable tool for precisely distinguishing between multiple closely related species of interest.

For assessing pathogenesis, all three assembly strategies enabled identification of genetic sequence associated with various adherence factors and regulatory elements within all the isolates. Differences were observed between methods due to statistical cut-offs for identity established prior to the analysis (Table 4). On average the MinION genome assemblies lacked the same degree of resolution and confidence in predicting the presence of several of the adherence factors resulting in several false-negatives. The MinION assemblies also appeared to possess multiple frameshifts and duplications, further complicating virulence factor analysis. The hybrid assemblies resulted in more accurate representation of the genomes of these bacteria.

The genome assemblies for the surrogate strains were scanned for the presence of gene sequences that encode virulence factor subunits (Table 4; details of virulence factors provided in Table S1). Although these lines were previously shown to lack functional virulence factors by other methods [34], the availability of these new complete genome assemblies enabled a more detailed investigation of the strains. Four strains (BAA-1427, BAA-1429, BAA-1430, and BAA-1431) possessed genetic sequences similar to those found within the enteropathogenic *E. coli* (EPEC) adherence factor plasmid (EAF) pB171. Sequences for bundle-forming pili (BFP) subunits BfpB (secretin), BfpE (inner membrane protein), and BfpH (transglycolase) [68] were identified in three MiSeq assemblies (BAA-1427, BAA-1430, and BAA-1431). Three of the hybrid and MinION counterparts (BAA 1427, BAA-1430, BAA 1431) lacked BfpH, and the MinION BAA-1430 lacked all three Bfp subunits. However, the noted absences following BLASTx analysis resulted from failure to meet the statistical threshold, likely reflecting nucleotide sequence variations. Additionally, all the assemblies for two strains (BAA-1429 and BAA-1430) encoded the BfpW/PerC transcriptional activators, which are part of the plasmid-encoded regulator (Per) responsible for BFP formation and activation of select genes within the locus of enterocyte effacement (LEE) [69–72]. Despite the presence of some subunits, these sequence elements are insufficient for formation of fully functional pili due to the absence of key accompanying subunit genes.

Apparent homologs of PapE (tip fimbriae), PapG (digalactoside-binding adhesion), and PapJ (assemble/integrity), were observed for all three assemblies of the BAA-1429 strain genome. These elements help comprise the pyelonephritis-associated pili/P fimbriae commonly seen in uropathogenic *E. coli* [73–76]. The pap operon is responsible for the formation of this pilus and has been previously described as encoding eleven distinct proteins (i.e., PapA, -B, -C, -D, -E, -F, -G, -H, -I, -J, -K). However, the presence of a fully functional P fimbriae pilus is unlikely due to the absence of fundamental structural and assembly elements [76–79].

Some form of a colonization factor (CF) that is typically observed in enterotoxigenic *E. coli* (ETEC) was observed in genome assemblies from all the strains. The most prevalent CF was the protein CsnA, which is a component of the major pilin monomer of the CS20 fimbriae [80–82]. However, in almost every instance, sequence similarity of these remained close to the statistical cut-off for identity, indicating lack of similarity to functional virulence factors. In addition to CFs, the BAA-1428 and BAA-1430 genomes contained genes similar to the CswA factor, commonly associated with the formation of the structural CS12 fimbriae subunits [81]. Lastly, *E. coli* BAA-1430 exclusively possessed sequence similarity to the CfaB, CooA, and CsbA CFs, associated with the colonization factor antigen I (CFA/I), CSI pilin major subunit, and the CS17 fimbrial subunit [81,83,84]. The CFs found within these genomes are associated with virulent ETEC; for pathogenesis to arise within these species there are two primary factors that must be present, which are the enterotoxins: heat-labile (LT) and/or heat-stable (ST) toxins, of which are frequently transported within the same plasmid as the CFs [81,84,85]. Their absence indicates these CFs may represent fimbriae/adherence factors that are not associated with pathogenesis, but confer similar structural properties to those that are. It is not uncommon that *E. coli* isolates, whether pathogenic or non-pathogenic, possess some mix of colonization and/or adherence factors [86–88]. However, in this experiment none of the identified sequences with resemblance to any putative virulence factor is expected to be functional or pose a hazard. The findings here indicate that the isolates of interest harbored no other detectable factors typically associated with virulent ETEC.

In two strain genomes (BAA-1427 and BAA-1430) potential homologous sequences encoding toxins associated with pathogenic *E. coli* were detected, but with only weak similarity -thus not likely functional. For example, the CNF1 holotoxin, an AB toxin, is documented to operate via RhoGTPases activity within eukaryotic cells but shared only 53.48% percent translated amino acid identity with the known functional toxin [89–91].

Documented regions within CNF1 that are responsible for host cell binding, as well the C-terminal portion that expresses the catalytic activity of this protein, were each ~50% identical in the surrogates when compared to the respective regions within their functional counterpart [92–94]. Additionally, the BAA-1427 and BAA-1430 genomes all appeared to harbor sequences sharing similarities with cytolethal distending toxin (CDT) [95,96]. CDT is a tripartite holotoxin responsible for cell cycle arrest and apoptosis within mammalian cells and is composed of a deoxyribonuclease-like toxin B-subunit and A and C subunits responsible for transporting the B subunit to the surface of the host cell [97–101]. Upon examination of these identified subunits, it was found that our samples only shared a 56.03% identity to subunit A (CAD48849.1), 69.87% to subunit B (CAD48859.1), and 40.56% to subunit C (CAD48851.1) when compared with functional toxins.

The availability of rifampicin-resistant strains derived from the surrogates also enabled screening for antibiotic resistant strains. All three sequencing and assembly methods enabled successful identification of SNPs within the *rpoB* gene known to confer rifampicin resistance [102,103]. As discussed, the genome assemblies generated from MinION data reflected the lowest quality assembly. While potentially useful for analysis of highly specific, known SNPS, these data would be difficult to utilize for discovery of new mutations (Table S2). Two isolates (BAA-1427 and BAA-1427 rif^R) appeared to share a silent mutation encoding amino acid A206, while the rif-mutant contained an additional L533 to P mutation, a change previously documented to confer rifampicin resistance [103]. Two strains (BAA-1428 and BAA-1428 rif^R) shared a silent mutation (T486 to T) compared to *E. coli* K-12. BAA-1428 rif^R also possessed an additional S512 to P mutation, which has not been previously reported to confer rifampicin resistance but resides within the first cluster of the rifampicin resistant determining region (RRDR) [102]. Lastly, the BAA-1430 genomes and BAA-1430 rif^R both shared silent mutations (P489 to P, L623 to L, and G846 to G), and the BAA-1430 rif^R possessed a H526 to Y mutation, also documented to confer drug resistance [103]. Ultimately, each technique was useful for identification of key differences between the parent surrogate strains and the K-12 reference *rpoB* gene, but the long-read genomes lacked the precision to accurately distinguish consistent key differences that conferred rifampicin resistance due to a high concentration of nucleotide deletions and false amino acid shifts.

5. Conclusions

This study provided a direct comparison of two common sequencing platforms and discussion of genome assembly characteristics applied to surrogate bacterial strains for which genome sequences were not available. From both research and regulatory standpoints, the application of WGS and subsequent bioinformatic analyses are indeed the tools of the future - unifying many traditionally used microbiological analyses into a singular workflow and enabling greater precision in surveillance of foodborne pathogens for quicker and more efficient regulatory response to foodborne outbreaks. Institutes such as the Center for Genomic Pathogen Surveillance have already adopted and standardized WGS-based pathogen detection. Similar research and application in both academe and industry will continue to accelerate. Both long- and short-read sequencing methods serve valuable, yet distinct roles for construction of complete microbial genomes. Long-read capacity facilitates better genome assembly and reveals structural properties of the genome that are not readily sequenced by other means. Short-read methods improve precision and resolution required for investigative studies and certain targeted analyses. As demonstrated in this study, when combined in hybrid fashion, the two sequencing approaches together are invaluable in enabling completed high-quality genomes to be constructed and accessible within databases for utilization in traceback and recalls in the instance of foodborne outbreaks in which a high degree of resolution is required for distinguishing between closely related bacterial strains. Libraries of high quality, complete genomes also serve a valuable research function, and provide a resource for further understanding of genomic sequences or alterations that can confer pathogenesis.

The application of true WGS (i.e., production of a complete genome comprised of a single contiguous sequence) as a means for daily routine screening within a food processing facility's food safety program is impractical and other forms of high-throughput sequencing may be more optimal. As demonstrated here, short-read platforms such as MiSeq provide a time- and cost-efficient means of simple of known pathogenic elements or antibiotic resistance genes. For the purposes of a food processing facility, confirmation of elements that confer pathogenesis is required, and while the methods described within this paper offered a means to achieve this, a fully closed bacterial genome is not required for most routine screening. While the cost of sequencing has greatly diminished as the quality of generated output continues to increase, WGS remains a data intensive process, relies on evaluation of DNA extracted from a pure bacterial culture, and is largely inefficient for routine screening. Ultimately, while routine WGS of samples taken within the food processing facility would serve a valuable means for differentiating what is being transported into the facility from the native microflora that pre-existed within the facility, it is impractical as a means for routine screening for outgoing lots. Although WGS enables sequence discrimination of genetic elements associated with both pathogenic and non-pathogenic strains, the current findings demonstrate that this outcome can be achieved more optimally via high-throughput targeted sequencing. Regardless, high-throughput sequencing methods will become increasingly important in food safety applications. These analyses enable insight into the genetic make-up of the surrogate strains studied that is useful for a variety of research applications and can also help inform decision-making for the incorporation and application of WGS within industry food safety programs.

Supplementary Materials: The following are available online at https://www.mdpi.com/2076-2607/9/3/608/s1, Figure S1: MinION Assemblies: Distribution of contig lengths. Figure S2: MiSeq Assemblies: Distribution of contig lengths. Table S1: List of virulence attributes examined. Table S2: Comprehensive list of alterations within the *rpoB* genes of each sample and assembly type.

Author Contributions: Conceptualization, H.R.C., G.C.S., T.M.T., P.K.R.; Data curation, K.K.; Formal analysis, D.A.T., K.K., J.J.G., B.W.D., J.M.; Funding acquisition, T.M.T., P.K.R.; Investigation, D.A.T., J.J.G., P.K.R.; Methodology, D.A.T., K.K., J.J.G., P.K.R.; Project administration, T.M.T., P.K.R.; Resources, T.M.T., P.K.R.; Software, K.K., B.W.D.; Supervision, H.R.C., T.M.T., P.K.R.; Validation, D.A.T., J.J.G., P.K.R.; Visualization, D.A.T.; Writing—original draft, D.A.T.; Writing—review and editing, D.A.T., K.K., J.J.G., B.W.D., A.E.H., J.M., H.R.C., G.C.S., T.M.T., P.K.R.. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the USDA Agriculture and Food Research Initiative Grant No. 2012-68003-30155 and the Texas Beef Council/Check-Off Program (T. M. Taylor), Beef Competitiveness Exceptional Item funding from the Texas legislature (P.K. Riggs), and Texas A&M AgriLife Research (T.M. Taylor and P.K. Riggs).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The completed genomes described within the text are available in a publicly accessible repository. NCBI GenBank accession numbers are located in Table 3.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Murano, A.E.; Cross, H.R.; Riggs, P.K. The outbreak that changed meat and poultry inspection system worldwide. *Anim. Front.* **2018**, *8*, 4–8. [CrossRef] [PubMed]
2. Allard, M.W.; Strain, E.; Melka, D.; Bunning, K.; Musser, S.M.; Brown, E.W.; Timme, R. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* **2016**, *54*, 1975–1983. [CrossRef]
3. Doyle, M.P.; Erickson, M.C.; Alali, W.; Cannon, J.; Deng, X.; Ortega, Y.; Smith, M.A.; Zhao, T. The food industry's current and future role in preventing microbial foodborne illness within the United States. *Clin. Infect. Dis.* **2015**, *61*, 252–259. [CrossRef] [PubMed]
4. Lüth, S.; Kelta, S.; Dahouk, S.A. Whole genome sequencing as a typing tool for foodborne pathogens like *Listeria monocytogenes*—The way towards global harmonization and data exchange. *Trends Food Sci. Tech.* **2018**, *73*, 67–75. [CrossRef]

5. Sekse, C.; Holst-Jensen, A.; Dobrindt, U.; Johannessen, G.S.; Li, W.; Spilsberg, B.; Shi, J. High throughput sequencing for detection of foodborne pathogens. *Front. Microbiol.* **2017**, *8*, 2029. [CrossRef]
6. Centers for Disease Control and Prevention (CDC). Burden of Foodborne Illnesses in the United States. Available online: http://www.cdc.gov/foodborneburden/burden/index.html (accessed on 4 April 2018).
7. Allard, M.W.; Bell, R.; Ferreira, C.M.; Gonzalez-Escalona, N.; Hoffmann, M.; Muruvanda, T.; Ottensen, A.; Ramachandran, P.; Reed, E.; Sharma, S.; et al. Genomics of foodborne pathogens for microbial food safety. *Curr. Opin. Biotechnol.* **2017**, *49*, 224–229. [CrossRef]
8. Jagadeesan, B.; Gerner-Smidt, P.; Allard, M.W.; Leuillet, S.; Winkler, A.; Xiao, Y.; Chaffron, S.; Vossen, J.V.D.; Tang, S.; Katase, M.; et al. The use of next generation sequencing for impriving food safety: Translation into practice. *Food Microbiol.* **2019**, *79*, 96–115. [CrossRef]
9. The National Human Genome Research Institute (NHGRI). Available online: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (accessed on 14 August 2019).
10. Taboada, E.N.; Graham, M.R.; Carrico, J.A.; Domselaar, G.V. Food safety in the age of next generation sequencing, bioinformatics, and open data access. *Front. Microbiol.* **2017**, *8*, 909. [CrossRef]
11. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351. [CrossRef] [PubMed]
12. Maio, N.D.; Shaw, L.P.; Hubbard, A.; George, S.; Sanderson, N.; Swann, J.; Wick, R.; AbuOun, M.; Stubberfield, E.; Hoosdally, S.J.; et al. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microb. Genom.* **2019**, *5*, 1–12. [CrossRef]
13. Ronholm, J.; Nasheri, N.; Petronella, N.; Pagotto, F. Navigating microbiological food safety in the era of whole-genome sequencing. *Clin. Microbiol. Rev.* **2016**, *29*, 837–856. [CrossRef] [PubMed]
14. Pollard, M.O.; Gurdasani, D.; Mentzer, A.J.; Porter, T.; Sandhu, M.S. Long reads: Their purpose and place. *Hum. Mol. Genet.* **2018**, *27*, R234–R241. [CrossRef] [PubMed]
15. Goodwin, S.; Gurtowski, J.; Ethe-Sayers, S.; Deshpande, P.; Schatz, M.C.; McCombie, W.R. Oxford nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **2015**, *25*, 1750–1756. [CrossRef]
16. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [CrossRef] [PubMed]
17. Loman, N.J.; Pallen, M.J. Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **2015**, *13*, 787–794. [CrossRef]
18. Deng, X.; Bakker, H.C.; Hendriksen, R.S. Genomic epidemiology: Whole-genome sequencing powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.* **2016**, *7*, 353–374. [CrossRef]
19. Jackson, B.R.; Tarr, C.; Strain, E.; Jackson, K.A.; Conrad, A.; Carleton, H.; Katz, L.S.; Stroika, S.; Gould, L.H.; Mody, R.K.; et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* **2016**, *63*, 380–386. [CrossRef] [PubMed]
20. Lakicevic, B.; Nastasijevic, I.; Dimitrijevic, M. Whole genome sequencing: An efficient approach to ensuring food safety. *IOP Conf. Ser. Earth Environ. Sci.* **2017**, *85*, 012052. [CrossRef]
21. Lytsy, B.; Engstrand, L.; Gustafsson, Å.; Kaden, R. Time to review the gold standard for genotyping vancomycin-resistant enterococci in epidemiology: Comparing whole-genome sequencing with PFGE and MLST in three suspected outbreaks in Sweden during 2013–2015. *Infect. Genet. Evol.* **2017**, *54*, 74–80. [CrossRef]
22. Rantsiou, K.; Kathariou, S.; Winkler, A.; Skandamis, P.; Saint-Cyr, M.J.; Rouzeau-Szynalski, K.; Amézquita, A. Next generation microbial risk assessment: Opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *Int. J. Food Microbiol.* **2018**, *287*, 3–9. [CrossRef]
23. Wang, S.; Daniel, W.; Falardeau, J.; Strawn, L.K.; Mardones, F.O.; Adell, A.D.; Switt, A.I.M. Food safety trends: From globalization of whole genome sequencing to application of new tools to prevent foodborne disease. *Trends Food Sci. Tech.* **2016**, *57*, 188–198. [CrossRef]
24. Portmann, A.; Fournier, C.; Gimonet, J.; Ngom-Bru, C.; Barretto, C.; Baert, L. A validation approach of an end-to-end whole genome sequencing workflow for source tracking of *Listeria monocytogenes* and *Salmonella enterica*. *Front. Microbiol.* **2018**, *9*, 446. [CrossRef]
25. Riggs, P.K. Expansion of knowledge and advances in genetics for quantitative analyses. In *Routledge Handbook of Sport and Exercise Systems Genetics*, 1st ed.; Lightfoot, J.T., Hubal, M.J., Roth, S.M., Eds.; Routledge: New York, NY, USA, 2019; pp. 16–27.
26. Cabrera-Diaz, E.; Moseley, T.M.; Lucia, L.M.; Dickson, J.S.; Castillo, A.; Acuff, G.R. Fluorescent protein-marked *Escherichia coli* biotype I strains as surrogates for enteric pathogens in validation of beef carcass interventions. *J. Food Prot.* **2009**, *72*, 295–303. [CrossRef]
27. Ingham, S.C.; Algino, R.J.; Ingham, B.H.; Schell, R.F. Identification of *Escherichia coli* O157:H7 surrogate organisms to evaluate beef carcass intervention treatment efficacy. *J. Food. Prot.* **2010**, *73*, 1864–1874. [CrossRef] [PubMed]
28. Keeling, C.; Niebuhr, S.E.; Acuff, G.R.; Dickson, J.S. Evaluation of *Escherichia coli* O157:H7 for cooking, fermentation, freezing, and refrigerated storage in meat processes. *J. Food Prot.* **2009**, *72*, 728–732. [CrossRef] [PubMed]

29. Marshall, K.M.; Niebuhur, S.E.; Acuff, G.R.; Lucia, L.M.; Dickson, J.S. Identification of *Escherichia coli* O157:H7 meat processing indicators for fresh meat through comparison of the effects of selected antimicrobial interventions. *J. Food. Prot.* **2005**, *68*, 2580–2586. [CrossRef] [PubMed]
30. Niebuhr, S.E.; Laury, A.; Acuff, G.R.; Dickson, J.S. Evaluation of nonpathogenic surrogate bacteria as process validation indicators for *Salmonella entericia* for selected antimicrobial treatments, cold storage, and fermentation in meat. *J. Food. Prot.* **2008**, *71*, 714–718. [CrossRef]
31. United States Department of Agriculture. Use of Non-Pathogenic *Escherichia coli* (*E. coli*) Cultures as Surrogate Indicator Organisms in Validation Studies. Available online: https://askfsis.custhelp.com/app/answers/detail/a_id/1392/~{}/use-of-non-pathogenic-escherichia-coli-%28e.-coli%29-cultures-as-surrogate (accessed on 10 November 2018).
32. American Type Culture Collection. Non-Pathogenic Escherichia coli Surrogate Indicators Panel. Available online: https://www.atcc.org/~{}/media/A0D8B646B84942088663F3B5A21CCAE8.ashx (accessed on 28 March 2019).
33. Dickson, J. (Iowa State University, Ames, IA, USA). Personal Communication, 2017.
34. American Type Culture Collection. Table 1: Summary of Virulence Testing Performed by the *E. coli* Reference Center of Pennsylvania State University (Provided by the Depositor). Available online: https://www.atcc.org/~{}/media/3C9579F7979A4968AE1F55A41360DC45.ashx (accessed on 28 March 2019).
35. Frenzel, M.A.; Savell, J.W.; Gehring, K.B.; Taylor, T.M.; Smith, G.C. Evaluation of Antimicrobial Interventions Applied during Further Processing of Raw Beef Products to Reduce Pathogen Contamination. Available online: https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/166061/FRENZEL-DISSERTATION-2017.pdf?sequence=1&isAllowed=y (accessed on 12 November 2018).
36. Kaspar, C.W.; Tamplin, M.L. Effects of temperature and salinity on the survival of *Vibrio vulnificus* in seawater and shellfish. *Appl. Environ. Microbiol.* **1993**, *59*, 2425–2429. [CrossRef] [PubMed]
37. Ausubel, F.M.; Brent, R.; Kingston, R.E.; Moore, D.D.; Seidman, J.G.; Smith, J.A.; Struhl, K. *Short Protocols in Molecular Biology*, 1st ed.; John Wiley & Sons: New York, NY, USA, 1989.
38. Chomczynski, P.; Mackey, K. Substitution of chloroform by bromochloropropane in the single-step method of RNA isolation. *Anal. Biochem.* **1995**, *1*, 163–164. [CrossRef]
39. Lannoy, C.; Riddler, D.; Risse, J. The long reads ahead: De novo genome assembly using the MinION [version 2; peer review: 2 approved]. *F1000 Res.* **2019**, *6*, 1083. [CrossRef]
40. Andrew, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://biostars.org/p/180392/ (accessed on 26 November 2018).
41. Bolger, M.A.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for illumine sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
42. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
43. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722–736. [CrossRef]
44. Quinlan, R.A.; Hall, I.M. Bedtools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
45. Wick, R.R.; Judd, L.M.; Gorrie, C.L.; Holt, K.E. Unicycler: Resolving bacterial genome assemblies from short and long read sequencing reads. *PLoS Comput. Biology* **2017**, *13*, e1005595. [CrossRef]
46. Walker, J.B.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Worman, J.; Young, S.K.; et al. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **2014**, *9*, e112963. [CrossRef] [PubMed]
47. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **2019**, *1962*, 227–245. [CrossRef]
48. Alonge, M.; Soyk, S.; Ramakrishnan, S.; Wang, X.; Goodwin, S.; Sedlazeck, F.J.; Lippman, Z.B.; Schatz, M.C. RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **2019**, *20*, 224. [CrossRef] [PubMed]
49. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
50. Joensen, K.G.; Tezchner, A.M.; Iguchi, A.; Aarestrup, F.M.; Scheutz, F. Rapid and easy in silico serotyping of *Escherichia coli* using whole genome sequencing (WGS) data. *J Clin. Microbiol.* **2015**, *53*, 2410–2426. [CrossRef] [PubMed]
51. Larsen, M.V.; Cosentino, S.; Rasmussen, S.; Friis, C.; Hasman, H.; Marvig, R.L.; Jelsback, L.; Sicheritz-Pontén, T.; Ussery, D.W.; Aarestrup, F.M.; et al. Multilocus sequence typing of total genome sequenced bacteria. *J. Clin. Mircobiol.* **2012**, *50*, 1355–1361. [CrossRef]
52. Chen, L.H.; Yang, J.; Yu, J.; Yao, Z.; Sun, L.; Shen, Y.; Jin, Q. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res.* **2005**, *33*, D325–D328. [CrossRef]
53. Chen, L.H.; Xiong, Z.H.; Sun, L.L.; Yang, J.; Jin, Q. VFDB 2012 update: Toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **2012**, *40*, D641–D645. [CrossRef]

54. Chen, L.H.; Zeng, D.D.; Lui, B.; Yang, J.; Jin, Q. VFDB 2016: Hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Res.* **2016**, *44*, D694–D697. [CrossRef] [PubMed]
55. Yang, J.; Chen, L.H.; Sun, L.L.; Yu, J.; Jin, Q. VFDB 2008 release: An enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.* **2008**, *36*, D539–D542. [CrossRef] [PubMed]
56. Afgan, E.; Baker, D.; Batut, B.; Beek, M.; Bouvier, D.; Cech, M.; Chilton, J.; Clements, D.; Coraor, N.; Gruning, B.; et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **2018**, *46*, W537–W544. [CrossRef]
57. Carver, T.; Harris, S.R.; Berriman, M.; Parkhill, J.; McQuillan, J.A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **2012**, *28*, 464–469. [CrossRef]
58. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
59. Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *K. Dan. Vidensk. Selsk.* **1948**, *5*, 1–34.
60. Blattner, R.F.; Plunkett, G.; Bloch, C.A.; Perna, N.T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J.D.; Rode, C.K.; Mayhew, G.F.; et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277*, 1453–1462. [CrossRef]
61. Benson, A.; Cavanaugh, M.; Clark, K.; Karsch-Mizranchi, I.; Lipman, D.J.; Ostell, J.; Sayers, E.W. GenBank. *Nucleic Acids Res.* **2013**, *41*, D36–D42. [CrossRef] [PubMed]
62. Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E.P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K.D.; Borodovsky, M.; Ostell, J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **2016**, *44*, 6614–6624. [CrossRef]
63. Janka, A.; Bielaszewska, M.; Dobrindt, U.; Greune, L.; Schmidt, M.A.; Karch, H. Cytolethal distending toxin gene cluster in enterohemorrhagic *Escherichia coli* O157:H7: Characterization and evolutionary considerations. *Infect. Immun.* **2003**, *71*, 3634–3638. [CrossRef] [PubMed]
64. Falbo, V.; Pace, T.; Picci, L.; Pizzi, E.; Caprioli, A. Isolation and nucleotide sequence of the gene encoding cytotoxic necrotizing factor 1 of *Escherichia coli*. *Infect. Immun.* **1993**, *61*, 4909–4914. [CrossRef] [PubMed]
65. Boža, V.; Brejová, B.; Vinař, T. Deepnano: Deep recurrent neural networks for base calling in MinION nanopore reads. *PLoS ONE* **2017**, *12*, e0178751. [CrossRef]
66. Tyler, D.A.; Mataseje, L.; Urfano, C.J.; Schmidt, L.; Antonation, K.S.; Mulvey, M.R.; Corbett, C.R. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Nat. Sci. Rep.* **2018**, *8*, 1–12. [CrossRef] [PubMed]
67. Wick, R.R.; Judd, L.M.; Gorrie, C.L.; Holt, K.E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genom.* **2017**, *3*, e000132. [CrossRef]
68. Tobe, T.; Hayashi, T.; Han, C.; Schoolnik, G.K.; Ohtsubo, E.; Sasakwaw, C. Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. *Infect. Immun.* **1999**, *67*, 5455–5462. [CrossRef]
69. Gómez-Duarte, O.G.; Kaper, J.B. A plasmid-encoded regulatory region activates chromosomal *eaeA* expression in enteropathogenic *Escherichia coli*. *Infect. Immun.* **1995**, *63*, 1767–1776. [CrossRef]
70. Mellies, J.L.; Elliot, S.J.; Sperandio, V.; Donnenberg, M.S.; Kaper, J.B. The Per regulon of enteropathogenic *Escherichia coli*: Identification of a regulatory cascade and a novel transcriptional activator, the locus of enterocyte effacement (LEE)-encoded regulator (Ler). *Mol. Microbiol.* **1999**, *33*, 296–306. [CrossRef] [PubMed]
71. Shin, S.; Castanie-Cornet, M.; Foster, J.W.; Crawford, J.A.; Brinkley, C.; Kaper, J.B. An activator of glutamate decarboxylase genes regulates the expression of enteropathogenic *Escherichia coli* virulence genes through control of the plasmid-encoded regulator, per. *Mol. Microbiol.* **2001**, *41*, 1133–1150. [CrossRef]
72. Tobe, T.; Schoolnik, G.K.; Sohel, I.; Bustamante, V.H.; Puente, J.L. Cloning and characterization of *bfpTVW*, genes required for the transcriptional activation of *bfpA* in enteropathogenic *Escherichia coli*. *Mol. Microbiol.* **1996**, *21*, 963–975. [CrossRef] [PubMed]
73. Kuehn, M.J.; Hueser, J.; Normark, S.; Hultgren, S.J. P pili in uropathogenic *E. coli* are composite fibres with distinct fibrillary adhesive tips. *Nature* **1992**, *356*, 252–255. [CrossRef] [PubMed]
74. Lane, M.C.; Mobley, H.L.T. Role of P-fimbrial-mediated adherence in pyelonephritis and persistence of uropathogenic *Escherichia coli* (UPEC) in the mammalian kidney. *Kidney Int.* **2007**, *72*, 19–25. [CrossRef]
75. Lillington, J.; Geibel, S.; Waksman, G. Biogenesis and adhesion of type 1 and P pili. *Biochim. Biophys. Acta.* **2014**, *1840*, 2783–2793. [CrossRef]
76. Tennent, J.M.; Lindberg, F.; Normark, S. Integrity of *Escherichia coli* P pili during biogenesis: Properties and role of PapJ. *Mol. Microbiol.* **1990**, *4*, 747–758. [CrossRef]
77. Goetz, G.S.; Mahmood, A.; Hultgren, S.J.; Engle, M.J.; Dodson, K.; Alpers, D.H. Binding of pili from uropathogenic Escherichia coli to membranes secreted by human colonocytes and enterocytes. *Infect. Immun.* **1999**, *67*, 6161–6163. [CrossRef] [PubMed]
78. Waksman, G.; Hultgren, S.J. Structural biology of the chaperone- usher pathway of pilus biogenesis. *Nat. Rev. Microbiol.* **2009**, *7*, 765–774. [CrossRef]
79. Wult, B.; Bergsten, G.; Samuelsson, M.; Svanborg, C. The role of p fimbriae for *Escherichia coli* establishment and mucosal inflammation in the human urinary tract. *Int. J. Antimicro. Agents* **2002**, *19*, 522–538. [CrossRef]
80. Mortezaei, N.; Epler, C.R.; Shao, P.P.; Shirdel, M.; Singh, B.; McVeigh, A.; Uhlin, B.E.; Savarino, S.J.; Andersson, M.; Bullitt, E. Structure and function of enterotoxigenic *Escherichia coli* fimbriae from differing assembly pathways. *Mol. Microbiol.* **2015**, *95*, 116–126. [CrossRef]

81. Nada, R.A.; Shaheen, H.I.; Khalil, S.B.; Mansour, A.; El-Sayed, N.; Touni, I.; Weiner, M.; Armstrong, A.W.; Klena, J.D. Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae. *J. Clin. Microbiol.* **2011**, *49*, 1403–1410. [CrossRef]
82. Valvatne, H.; Steinsland, H.; Grewal, H.M.S.; Mølbak, K.; Vuust, J.; Sommerfelt, H. Identification and molecular characterization of the gene encoding coli surface antigen 20 of enterotoxigenic *Escherichia coli*. *FEMS MicrobioMicrobiol. Lett.* **2004**, *239*, 131–138. [CrossRef]
83. Galkin, V.E.; Kolappan, S.; Ng, D.; Zong, Z.; Li, J.; Yu, X.; Egelman, E.H.; Craig, L. The structure of the CS1 pilus of enterotoxigenic *Escherichia coli* reveals structural polymorphisms. *J. Bacteriol.* **2013**, *195*, 1360–1370. [CrossRef] [PubMed]
84. Nataro, J.P.; Kaper, J.B. Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* **1998**, *11*, 142–201. [CrossRef] [PubMed]
85. Wiley, J.M.; Sherwood, L.M.; Woolverton, C.J. *Prescott's Microbiology*, 8th ed.; McGraw-Hill Higher Education: New York, NY, USA, 2010.
86. Frömmel, U.; Böhm, A.; Nitschke, J.; Weinreich, J.; Grob, J.; Rödiger, S.; Wex, T.; Ansorge, H.; Zinke, O.; Schröder, C.; et al. Adhesion patterns of commensal and pathogenic *Escherichia coli* from humans and wild animals on human and porcine epithelial cell lines. *Gut Pathog.* **2013**, *5*, 31. [CrossRef]
87. Grozdanov, L.; Raasch, C.; Schulze, J.; Sonnenborn, U.; Gottschalk, G.; Hacker, J.; Dobrindt, U. Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917. *J. Bacteriol.* **2004**, *186*, 5432–5441. [CrossRef]
88. Rendón, A.M.; Saldaña, Z.; Erdem, A.L.; Monteiro-Neto, V.; Vázquez, A.; Kaper, J.B.; Puente, J.L.; Girón, J.A. Commensal and pathogenic Escherichia coli use a common pilus adherence factor for epithelial cell colonization. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 10637–10642. [CrossRef]
89. Caprioli, A.; Falbo, V.; Ruggeri, M.; Baldassarri, L.; Bisicchia, R.; Ippolito, G.; Romoli, E.; Donelli, G. Cytotoxic necrotizing factor production by hemolytic strains of *Escherichia coli* causing extraintestinal infections. *J. Clin. Microbiol.* **1987**, *25*, 146–149. [CrossRef] [PubMed]
90. Khan, N.A.; Wang, Y.; Kim, K.J.; Chung, J.W.; Wass, C.A.; Kim, K.S. Cytotoxic necrotizing factor-1 contributes to *Escherichia coli* K1 invasion of the central nervous system. *J. Biol. Chem.* **2002**, *277*, 15607–15612. [CrossRef] [PubMed]
91. Knust, Z.; Blumenthal, B.; Aktories, K.; Schimdt, G. Cleavage of *Escherichia coli* cytotoxic necrotizing factor 1 is required for full biological activity. *Infect. Immun.* **2009**, *77*, 1835–1841. [CrossRef]
92. Fabbri, A.; Gauthier, M.; Bouquet, P. The 5′ region of *Cnf1* harbours a translational regulatory mechanism for CNF1 synthesis and encodes the cell-binding domain of the toxin. *Mol. Microbiol.* **1999**, *33*, 108–118. [CrossRef] [PubMed]
93. Fabbri, A.; Travaglione, S.; Fiorentini, C. Escherichia coli cytotoxic necrotizing factor 1 (cnf1): Toxin biology, in vivo applications and therapeutic potential. *Toxins* **2010**, *2*, 283–296. [CrossRef]
94. Lemichez, E.; Flatau, G.; Bruzzone, M.; Boquet, P.; Gauthier, M. Molecular localization of the *Escherichia coli* cytotoxic necrotizing factor cnf1 cell-binding and catalytic domains. *Mol. Microbiol.* **1997**, *24*, 1061–1070. [CrossRef]
95. Friedrich, A.W.; Lu, S.; Bielaszewska, M.; Prager, R.; Bruns, P.; Xu, J.; Tschäpe, H.; Karch, H. Cytolethal distending toxin in *Escherichia coli* O157:H7: Spectrum of conservation, structure, and endothelial toxicity. *J. Clin. Microbiol.* **2006**, *44*, 1844–1846. [CrossRef]
96. Guerra, L.; Cortes-Bratti, X.; Guidi, R.; Frisan, T. The biology of the cytolethal distending toxins. *Toxins* **2011**, *3*, 172–190. [CrossRef] [PubMed]
97. Ceelen, L.M.; Decostere, A.; Ductatelle, R.; Haesebrouck, F. Cytolethal distending toxin generates cell death by inducing a bottleneck in the cell cycle. *Microbiol. Res.* **2006**, *161*, 109–120. [CrossRef]
98. Elwell, C.; Chao, K.; Patel, K.; Dreyfus, L. *Escherichia coli* cdtb mediates cytolethal distending toxin cell cycle arrest. *Infect. Immun.* **2001**, *69*, 3418–3422. [CrossRef]
99. Nesic, D.; Hsu, Y.; Stebbins, C.E. Assembly and function of a bacterial genotoxin. *Nature* **2004**, *429*, 429–433. [CrossRef] [PubMed]
100. Pickett, C.L.; Whitehouse, C.A. The cytolethal distending toxin family. *Nat. Rev.* **1999**, *7*, 292–297. [CrossRef]
101. Tejero, M.L.; Galán, J.E. Cdta, cdtb, and cdtc form a tripartite complex that is required for cytolethal distending toxin activity. *Infect. Immun.* **2001**, *69*, 4358–4365. [CrossRef]
102. Goldstein, P.B. Resistance to rifampicin: A review. *J. Antibiot.* **2014**, *67*, 625–630. [CrossRef] [PubMed]
103. Jin, J.D.; Gross, C.A. Mapping and sequencing of mutations in the *Escherichia coli rpoB* gene that lead to rifampicin resistance. *J. Mol. Biol.* **1988**, *202*, 45–58. [CrossRef]