Comparative genomics reveals structural and functional features specific to the genome of a foodborne Escherichia coli O157:H7

Vijay K. Sharma 1*, Suryatej Akavaram 1, Robert G. Schaut 1,3 and Darrell O. Bayles 2

Abstract

Background: Escherichia coli O157:H7 (O157) has been linked to numerous foodborne disease outbreaks. The ability to rapidly sequence and analyze genomes is important for understanding epidemiology, virulence, survival, and evolution of outbreak strains. In the current study, we performed comparative genomics to determine structural and functional features of the genome of a foodborne O157 isolate NADC 6564 and infer its evolutionary relationship to other O157 strains.

Results: The chromosome of NADC 6564 contained 5466 kb compared to reference strains Sakai (5498 kb) and EDL933 (5547 kb) and shared 41 of its 43 Linear Conserved Blocks (LCB) with the reference strains. However, 18 of 41 LCB had inverse orientation in NADC 6564 compared to the reference strains. NADC 6564 shared 18 of 19 bacteriophages with reference strains except that the chromosomal positioning of some of the phages differed among these strains. The additional phage (P19) of NADC 6564 was located on a 39-kb insertion element (IE) encoding several hypothetical proteins, an integrase, transposases, transcriptional regulators, an adhesin, and a phosphoethanolamine transferase (PEA). The complete homologs of the 39-kb IE were found in E. coli PCN061 of porcine origin. The IE-encoded PEA showed low homology (32–33%) to four other PEA in NADC 6564 and PEA linked to mobilizable colistin resistance in E. coli but was highly homologous (95%) to a PEA of uropathogenic, avian pathogenic, and enteroaggregative E. coli. NADC 6564 showed slightly higher minimum inhibitory concentration of colistin compared to the reference strains. The 39-kb IE also contained dndBCDE and dptFGH operons encoding DNA S-modification and a restriction pathway, linked to oxidative stress tolerance and self-defense against foreign DNA, respectively. Evolutionary tree analysis grouped NADC 6564 with lineage I O157 strains.

Conclusions: These results indicated that differential phage counts and different chromosomal positioning of many bacteriophages and genomic islands might have resulted in recombination events causing altered chromosomal organization in NADC 6564. Evolutionary analysis grouped NADC 6564 with lineage I strains and suggested its earlier divergence from these strains. The ability to perform S-DNA modification might affect tolerance of NADC 6564 to various stressors.

Keywords: O157, Genomics, Bacteriophages, Mobile elements, Genomic islands, Stress response
Background

Enterohemorrhagic *Escherichia coli* O157:H7 (O157) is a zoonotic human pathogen, transmitted through the consumption of contaminated foods, such as beef and dairy products, ready-to-eat salad greens, vegetables, and fruits [1–5]. O157 infections are the predominant cause of hemorrhagic uremic syndrome and kidney failure, especially in children and elderly [6, 7]. Cattle are the primary reservoir of O157 and are asymptomatic carriers of these bacteria in their gastrointestinal tract [8]. Carrier cattle shed O157 in their feces, which is a major risk factor in the contamination of meats produced from these animals [4, 9–11]. Since the first reported disease outbreak linked to O157 in the early 1980s [12], numerous outbreaks implicating O157 have occurred in the USA [13].

A variety of approaches, such as pulse-field gel electrophoresis [14, 15], multilocus genotyping [16–19], PCR- or microarray-based DNA fingerprinting [20–22], single nucleotide polymorphism (SNP) analysis [23, 24], and whole genome sequencing have been used to characterize outbreak strains and to understand their epidemiology, virulence, and evolutionary relatedness [25–28]. Some of these approaches have allowed grouping of O157 isolates into lineages I, I/II, and II. The isolates in lineages I and I/II are predominantly recovered from infected humans and those in lineage II are predominantly cattle-associated [21, 29]. Several lineage I isolates that have been linked to large disease outbreaks possess genetic alterations that cause increased expression of virulence genes, thus enhancing morbidity and the likelihood of serious disease in infected individuals [27, 30, 31]. In addition, genetic variability caused by single nucleotide polymorphisms or insertion and deletions in certain loci are known to enhance expression of genes linked to biofilm formation among O157 isolates [32–36]. It has been postulated that O157 isolates exist as heterogeneous populations capable of expressing genes for colonizing biotic surfaces, such as gastrointestinal tracts of carrier animals and incidental human host as well as surfaces of leafy vegetables, and abiotic surfaces through the formation of biofilms [29, 31–34, 37–44].

Whole genome sequencing of several O157 isolates has revealed that horizontally-acquired DNA, called O islands, in the core of the *E. coli* genome has contributed to the emergence of pathogenic O157 strains from a non-pathogenic ancestral *E. coli* [25, 45, 46]. Many of these O islands represent mobile elements, such as bacteriophages or insertion elements, encoding functions related to virulence [25, 47]. Prominent among these are the locus of enterocyte effacement (LEE) and genes encoding for Shiga toxins [25, 48]. While LEE promotes intimate adherence of O157 to intestinal epithelial cells [49, 50], Shiga toxins damage microvasculature of intestinal mucosa and kidneys to produce hemorrhagic colitis and hemorrhagic uremic syndrome, respectively [51].

According to a recent phylogenetic analysis, O157 strains once shared a common ancestor with enteropathogenic *E. coli* (EPEC) O55:H7 and then diverged from EPEC [46]. This divergence is characterized by the acquisition of plasmid pO157, bacteriophages encoding Shiga toxins 1 and 2, a novel O antigen-encoding gene cluster, and by loss of functions, such as the ability to ferment sorbitol and express beta-glucuronidase [52, 53]. A subsequent report estimated that the average mutation rate is about 50% faster in the O157 lineage compared to the O55:H7 lineage, suggesting a more recent divergence of these two lineages as opposed to an earlier divergence time point [46].

Besides acquiring virulence genes through horizontal gene transfer, bacterial pathogens may also acquire genomic islands encoding genes for survival under multiple stresses. The *dnd* operon, originally identified in *Streptomyces lividans* 66, encodes enzymes for sequence-specific, post-replication swapping of non-bridging oxygen with a sulfur in the DNA backbone [54, 55]. This DNA S-modification has been shown to enable bacterial growth under a variety of stressful conditions, such as extreme temperature, pH, UV, X-ray, salinity, pressure and heavy metals, by protecting genomic DNA and proteins from intracellular oxidative damage [56]. Not all, but many bacterial species harboring *dnd* genes also carry an upstream, three-gene operon (*dptFGH*) encoding a DNA S-modification-dependent restriction system for restricting heterologous DNA and protecting the host S-modified DNA [57]. The genes in the *dnd* and *dpt* operons, which have been identified in phylogenetically diverse bacterial species, including both non-pathogenic and pathogenic *E. coli* [58], exhibit a high degree of synteny. The presence of these operons on genomic islands or mobile elements indicates that these genes spread from an ancestral species to other bacterial species during the course of evolution [59].

Similarly, acquisition of antibiotic resistance genes via mobile DNA elements enhances bacterial survival in the presence of antibiotics. Colistin has been a last-resort treatment option for multidrug-resistant gram-negative pathogens, including members of Enterobacteriaceae; however, the emergence of colistin resistance had rendered this antibiotic ineffective. One of the mechanisms enhancing resistance to colistin involves phosphoethanolamine substitution in the outer membrane lipid A that reduces the ability of colistin to enter and kill bacterial cells [60]. The enzymes phosphoethanolamine transferases (PEA) constitute a family of related proteins that are responsible for lipid A modification. Some of these PEA have evolved to confer a high level of mobilizable colistin resistance (MCR), which is encoded by the *mcr* genes present on mobile elements that are capable of transmission to other gram-negative bacterial species [60, 61]. According to a recent report, O157 carries a chromosomal PEA gene *pmrC* and several other PEA genes mediating lipid A modification that contribute to a
Results

General features of strain NADC 6564 chromosome
The complete genome of NADC 6564 was sequenced using a combination of PacBio RS II and Illumina MiSeq sequencing methods and assembled and annotated into a chromosome of 5466,770 bp and a plasmid of 92,691 bp as described previously [28]. In the current study, we used BLAST Ring Image Generator (BRIG) to generate a detailed circular map of the annotated NADC 6564 chromosomal sequence (Fig. 1a). This chromosomal map shows predicted protein-coding sequences (CDS) from both its positive and negative strands, which are represented by the two outermost rings (Fig. 1a). These CDS are shown as purple or yellow rectangles, where the size of each rectangle corresponds to the length of the CDS. A total of 19 bacteriophages (labeled P1 – P19) identified in the assembled sequence of NADC 6564 are shown as pink rectangles outside of the positive-chromosomal strand (Fig. 1a). The size of the chromosome, number of CDS, and number of genes encoding rRNA and tRNA were very similar between NADC 6564 and the four O157 reference strains (Table 1). The chromosome of NADC 6564 contained 53 genomic islands (GI) compared to reference strains EDL933, Sakai, and TW14359, containing 63, 71 and 44 GI, respectively (Table 1).

Altered organization of conserved genomic blocks in the chromosome of NADC 6564

On the average, the chromosome of O157 strains are about 1.40 Mb larger than that of non-pathogenic E. coli K12 strains [47]. All E. coli strains share a common core genome of about 4.1 Mb organized into linear conserved blocks (LCB) that are disrupted in pathogenic E. coli lineages by the integration of laterally-acquired genetic elements, such as genomic islands (GI) and bacteriophages [47]. These genetic elements carry a wide variety of genes impacting bacterial virulence, metabolism, and genetic organization of chromosomal DNA [25, 27, 72]. In order to determine the basic arrangement and organization of LCB, we compared the Mauve alignment of the whole chromosomal sequence of NADC 6564 to reference O157, and several non-O157 Shiga toxin-producing E. coli (STEC) strains. The results of this comparative alignment showed that the chromosomes of compared strains were organized into 40 to 42 LCB (Fig. 1b). These LCB were arranged contiguously from left to right in all compared strains. However, 18 LCB (numbered 5–22 when counting from the left) of strain NADC 6564, which are underlined with a long black arrow, were arranged in an opposite order compared to that in other O157 and non-O157 strains (Fig. 1b). The LCB 18–19 (identified by a short black arrow) had the same orientation in strains NADC 6564 and Sakai compared to that in EDL933 (Fig. 1b).

Dissimilar types, number, and distribution of bacteriophages and genomic islands in NADC 6564 compared to reference strains

Since arrangement of 18 out of 41 LCB was different in NADC 6564, we wanted to analyze location of laterally acquired genetic elements, such as bacteriophages and genomic islands in NADC 6564 relative to the reference strains. Moreover, the understanding of differences in the chromosomal location of diverse sets of bacteriophages and mobile elements could provide insight into the mechanisms by which these elements differentially impact virulence and other traits of O157 strains [72]. Analysis of the chromosomal sequences by PHASTER [73, 74], a software for identifying phage-like elements in bacterial genomes, resulted in the identification of 19 (P1 – P19) such elements in NADC 6564 (Table 2 and Fig. 2) compared to 18 (P1 – P18) such elements identified in the chromosomes of reference strains EDL933 and Sakai (Additional file 1: Table S2). The locations of these phages are shown outside of the outermost ring of the comparative circular chromosomal map of NADC 6564 and the reference strains created by BRIG [75] (Fig. 3a). The 19 phages of str. NADC 6564 were
inserted in different LCB at sites that were either identical or different from the insertion sites occupied by the corresponding similar phages in two reference strains (Fig. 2, Table 2, and Additional file 1: Table S2). For example, phage 933 W, which carries the stx2 gene, occupied the same insertion site that was flanked by the gene (wbrA) encoding NAD(P)H: quinone oxidoreductase in str. NADC 6564 and the two reference strains (EDL933 and Sakai). About 10 of the 19 phages in NADC 6564 and the reference strains EDL933 and Sakai are inserted in the region of the genome containing LCB 5–22. The remaining phages were inserted in LCB that are located immediately surrounding LCB 5–22 (Fig. 2 and Table 2). We also analyzed NADC 6564 chromosomal sequence...
to identify genomic islands (GI) and their locations on this chromosome and compare the results of this analysis with similarly analyzed chromosomal sequences of reference strains. Based on this analysis, we identified 53 GI in the chromosome of NADC 6564 compared to the presence of 63 and 71 GI in the reference strains EDL933 and Sakai, respectively (Additional file 2: Table S3 and Fig. 3a). The overall representation of GI amounted to about 17.08% of the total chromosomal sequence in NADC 6564 compared to about 19.37 and 15.16% for the GI in strains EDL933 and Sakai, respectively. Most of the GI occupied the same LCB in the chromosome of NADC 6564 and the reference strains that had phages integrated in them (Additional file 2: Table S3 and Fig. 3a). This preferential distribution of GI and phages in this chromosomal region (approximately 1000 kb – 3500 kb) indicates a propensity of this region for increased DNA recombination and rearrangements resulting in differential arrangements of LCB in O157 strains.

### Relationship of strain NADC 6564 to lineage I versus the other lineages

O157 is considered an emerging foodborne pathogen presumably due to reported genetic heterogeneity of O157 populations enabling them to diverge into isolates that differ in their abilities to colonize the carrier host animal, cause disease in humans, or survive in the external environment [33–35, 76]. Population heterogeneity is driven by intrinsic genetic changes, such as point mutations, small DNA insertions or deletions, as well as through the lateral acquisition of genomic islands and bacteriophages [25, 30, 72]. Since genome heterogeneity

### Table 1 Comparison of basic features of the strain NADC 6564 chromosome and reference O157:H7 strains

| Strains  | Size (Mb) | CDS | tRNA genes | rRNA genes | Genomic Islands (GI) | Size range GI (kb) | % GC | Reference | Source |
|----------|-----------|-----|------------|------------|----------------------|-------------------|------|-----------|--------|
| NADC 6564 | 5.46      | 5542 | 103        | 22         | 53                   | 4–85              | 50.0 | [28]      | Ground Beef |
| EDL933   | 5.53      | 5587 | 100        | 22         | 63                   | 3–88              | 50.4 | [80]      | Ground Beef |
| Sakai    | 5.49      | 5504 | 103        | 22         | 71                   | 4–59              | 50.5 | [72]      | Sprouts    |
| TW14359  | 5.52      | 5555 | 108        | 22         | 44                   | ND                | 50.5 | [27]      | Spinach    |
| EC4115   | 5.57      | 5608 | 109        | 22         | ND                   | ND                | 50.5 | [25]      | Spinach    |

*Accession numbers: CP017251.1 (NADC 6564), CP008957.1 (EDL933), BA000007.2 (Sakai), CP001368.1 (TW14359), and NC_011353.1 (EC4115); CDS Protein-coding sequence, ND Not determined

### Table 2 Bacteriophage profile of strain NADC 6564

| Bacteriophage Region/Name | Sequence Length (kb) | Number of Phage Proteins | Chromosomal Location | Genes Flankinga Phage Insertion Site | GC% | Accession Number |
|---------------------------|----------------------|--------------------------|----------------------|-------------------------------------|------|------------------|
| 1/Stx2-converting phage 1717 | 21                   | 7                        | 1,172,843–1,193,884 | BHW77_05970 and BHW77_06120     | 41.96 | NC_011357  |
| 2/Shigella phage Sf6       | 6.4                  | 4                        | 1,473,190–1,479,623 | BHW77_07465 and tRNA-Arg           | 46.18 | NC_025344  |
| 3/Enterobacteria phage YYZ-2008 | 32.1                | 18                       | 1,723,923–1,756,041 | BHW77_09510 and BHW77_09560      | 51.91 | NC_004813  |
| 4/Enterobacteria phage BP-4795 | 6.8                  | 3                        | 1,899,932–1,906,773 | BHW77_09625 and BHW77_09965      | 51.37 | NC_004813  |
| 5/Enterobacteria phage BP-4795 | 56.5                 | 14                       | 1,922,013–1,978,603 | BHW77_10140 and BHW77_10630      | 50.16 | NC_000924  |
| 6/Enterobacteria phage 933 W | 70.2                 | 65                       | 2,015,661–2,085,889 | BHW77_11820 and BHW77_12200      | 49.29 | NC_004813  |
| 7/Enterobacteria phage BP-4795 | 53.1                 | 24                       | 2,308,130–2,361,308 | BHW77_19495 and BHW77_19800      | 50.11 | NC_009514  |
| 8/Brucella phage P88       | 15.2                 | 18                       | 2,364,182–2,379,410 | BHW77_15525 and BHW77_15920      | 50.15 | NC_004813  |
| 9/Escherichia virus Lambda | 53.3                 | 9                        | 2,385,028–2,438,396 | BHW77_15055 and BHW77_15384      | 50.30 | NC_011357  |
| 10/Stx2-converting phage 1717 | 2.6                  | 3                        | 2,428,442–2,431,075 | BHW77_12210 and BHW77_12335      | 51.17 | NC_011357  |
| 11/Enterobacteria phage YYZ-2008 | 113.4                | 31                       | 2,514,193–2,627,682 | BHW77_17080 and BHW77_17080      | 50.26 | NC_004813  |
| 12/Enterobacteria phage BP-4795 | 60                   | 29                       | 2,800,143–2,860,221 | BHW77_17080 and BHW77_17080      | 50.25 | NC_004813  |
| 13/Enterobacteria phage BP-4795 | 57.4                 | 32                       | 2,967,268–3,024,670 | BHW77_17080 and BHW77_17080      | 50.94 | NC_004813  |
| 14/Enterobacteria phage P88 | 24.8                 | 18                       | 3,361,891–3,386,762 | BHW77_17080 and BHW77_17080      | 50.31 | NC_026014  |
| 15/Escherichia virus Lambda | 30.1                 | 14                       | 3,440,713–3,470,640 | BHW77_17080 and BHW77_17080      | 54.39 | NC_004813  |
| 16/Shigella phage POCJ13   | 24.1                 | 6                        | 3,463,508–3,487,685 | BHW77_17080 and BHW77_17080      | 47.40 | NC_025343  |
| 17/Enterobacteria phage cdtI | 48.2                 | 22                       | 3,709,593–3,757,796 | BHW77_17080 and BHW77_17080      | 50.11 | NC_009514  |
| 18/Enterobacteria phage Sf1 | 26.1                 | 6                        | 4,330,960–4,357,158 | BHW77_17080 and BHW77_17080      | 45.57 | NC_027339  |
| 19/Acidianus tailed spindle virus | 10.1                 | 1                        | 4,801,573–4,811,702 | BHW77_17080 and BHW77_17080      | 46.57 | NC_029316  |

aGenes flanking the phage insertion sites are represented as locus-tags in the annotated sequence (Accession number: CP017251.1)
has been used for classifying O157 strains into three major evolutionary lineages I, I/II, and II [21, 45], we performed comparative BLAST analysis to classify the lineage of NADC 6564. Figure 3a summarizes the comparative BLAST homologies of NADC 6564 to O157 strains of lineage I, II, and I/II, to non-O157 Shiga toxin-producing E. coli strains, and to a non-pathogenic E. coli K12 strain MG1655. As shown in Fig. 3a, greatest homology was observed between the chromosomal sequence of NADC 6564 and the lineage I O157 strains (EDL933 and Sakai) because only a few regions (4–5), each consisting of a short stretches of nucleotides, were missing in strains EDL933 and Sakai compared to NADC 6564.

The chromosomes of lineages II (FRIK2533) and I/II (TW14359) O157 strains were lacking 24 to 26 regions of variable lengths that were present in NADC 6564 (Fig. 3a). The non-O157 Shiga toxin-producing E. coli K12 strain MG1655 and E. coli O26 and O111 strains had the highest number of regions missing in their chromosomes compared to NADC 6564 and other O157 strains irrespective of their lineage affiliation (Fig. 3a).

BLAST results identified a novel 39-kb insertion element in NADC 6564

Despite extensive whole chromosome sequence homology between NADC 6564 and other O157 strains of lineage I, as revealed by BLAST analysis, a 39-kb (4,789,473 bp – 4,789,529 bp) region carrying bacteriophage P19 was present only in the chromosome of NADC 6564 and was missing from chromosomes of all other O157 and non-O157 strains (Fig. 3b). This 39-kb insertion element (IE) is bounded by a 56-bp direct repeat at its 5’ (4,789,473 bp – 4,789,529 bp) and 3’ (4828,590 bp – 4828,646 bp) ends and is inserted near the gene encoding for tRNA-Leu. BLAST analysis confirmed that none of the O157 and non-O157 strains, whose chromosomal sequences were available in the GenBank, contained this 39-kb IE. However, this IE was present in the unpublished genomes of E. coli strain AR0015 (Accession Number: CP024862.1) and E. coli strain C1 (Accession Number: CP010116.1). The genome of a porcine E. coli strain PCN061 contained two regions, one 23 kb and the other 7.1 kb in size, that were homologous to the 39-kb IE with only 1.0 kb non-homologous DNA separating the two homologous regions [77].

The 39-kb IE encodes genes for mobility, potential virulence, and stress tolerance in strain NADC 6564

As shown in Fig. 3b, the 39-kb IE of strain NADC 6564 encodes for an integrase located near its 5’ end, which is inserted near the tRNA-Leu gene. The 3’ region of this IE also encodes a transposase (in addition to those encoded on P19), GTPase, and a transcriptional regulator in the 3’ region. The presence of genes encoding an integrase and a transposase, and the presence of a 56-bp direct repeat at the ends, are all suggestive that the 39-kb DNA region is an insertion element acquired independently by NADC 6564. In addition, the 39-kb IE also contains genes encoding for a phosphoethanolamine transferase, adhesion, and dptFGH and dndBCDE operons predicted to encode DNA phosphorothioation or (S-DNA modification) and a restriction system (Table 3) for recognizing and degrading
DNA lacking the S-DNA modification, respectively [57, 58]. However, the dndA gene, which represents the fifth gene of the dnd operon (dndABCDE) in many bacterial species [59], is lacking in the genome of NADC 6564. In *E. coli* and many other bacterial species that contain only four- (dndBCDE) instead of five-gene dnd operon (dndABCDE), the iscS gene represents an ortholog of the dndA gene [78, 79]. IscS functions as a cysteine desulfurase which is a key enzyme of DNA phosphorothioation pathway [79]. The NADC
virulence of these strains has not been fully described. The genomic DNA of O157 strains EDL933 and Sakai, and these two strains was susceptible to degradation by an iodine solution compared to higher number of colonies recovered from plates that were surface-spread with NADC 6564 transformed with S-DNA-modified pUC19 (recovered from pUC19-transformed NADC 6564) or transforming NADC 6564 with S-DNA-modified pUC19 (recovered from pUC19-transformed NADC 6564) or transforming NADC 6564 dptFGH deletion mutant and EDL933 with pUC19 would show equivalent recovery of the transformed DNA, respectively, to rule out differences in the transformation efficiency of these two strains. Although the dndBCDE and dptFGH operons are widely distributed among diverse groups of bacterial species, including pathogenic and non-pathogenic E. coli [56–58], BLAST analysis of dndBCDE (4,808,986 bp – 4,814,048 bp; Accession Number CP017251.1) and dptFGH (4,815,824 bp – 4,823,838 bp, Accession Number CP017251.1) sequences against the nucleotide sequences in GenBank revealed that NADC 6564 did not retrieve any O157 strains carrying these genes. However, sequences showing 99% homology to the DNA regions carrying dndBCDE and dptFGH operons of NADC 6564 were identified in a porcine E. coli strain PCN061 [77] and two other unpublished genomes of E. coli strains (AR0015: Accession Number CP024862.1; AR436: Accession Number CP029111.1). The sequences exhibiting homology ranging from 78 to 81% to nucleotide sequences of the dndBCDE operons by other bacterial species in Enterobacteriaceae produced multiple clades of NADC 6564 were also identified in Salmonella enterica (Accession Number: CP019412.1), Enterobacter sp. 638 (Accession Number: CP000653.1), Yersinia ruckeri str. YRB (Accession Number: CP009539.1), and Erwinia amylovora str. E2 (Accession Number: CP024970.1). An evolutionary tree generated by comparing the homology of genes in the dndBCDE (Additional file 3: Figure S1) and dptFGH (Additional file 4: Figure S2) operons to bacterial species in Enterobacteriaceae produced multiple clades over the course of the evolution of these bacterial species.

6564 does encode iscS which occupies 1,289,336–1,290,550 bp in its genome. Although genes showing 100% homology to the adhesin gene encoded in the 39-kb IE of NADC 6564 are present in extraintestinal E. coli strain PCN061 [77], avian pathogenic E. coli strain O1 (GenBank Accession Number: CP028310.1), uropathogenic E. coli strain UT189 (GenBank Accession Number: CP000243.1), and enterohemorrhagic E. coli strain 104:H21 [80], the contribution of this adhesin to virulence of these strains has not been fully described.

### The dndBCDE operon conferred S-DNA modification that protected modified DNA from restriction activity encoded by the dptFGH operon

The function of the dndBCDE operon encoded by the 39-kb IE in NADC 6564 and NADC 6565 (a Congo red–positive variant of NADC 6564 included as a control) [28] was verified by demonstrating that the genomic DNA of these two strains was susceptible to degradation by an iodine solution compared to no effect of this treatment on the genomic DNA of O157 strains EDL933 and Sakai, and non-pathogenic E. coli TOP10 lacking this operon (Fig. 4a). These results indicated that genes in the dndBCDE operon are expressed and their encoded proteins are able to S-modify DNA in NADC 6564. Similarly, we also demonstrated that the presence of dptFGH operon resulted in significantly (p < 0.05) lower recovery of colonies on agar plates that were surface-spread with NADC 6564 transformed with plasmid pUC19 DNA lacking S-modification compared to higher number of colonies recovered from O157 EDL933 transformed with same plasmid DNA (Fig. 4b and c). These results indicated that the dptFGH operon in NADC 6564 encodes for a restriction system capable of restricting incoming heterologous DNA (pUC19) devoid of S-DNA modification compared to strain EDL933 lacking the dptFGH operon. However, we did not investigate whether transforming strains NADC 6564 and EDL933 with S-DNA-modified pUC19 (recovered from pUC19-transformed NADC 6564) or transforming NADC 6564 dptFGH deletion mutant and EDL933 with pUC19 would show equivalent recovery of the transformed DNA.

### Phosphoethanolamine transferase encoded on the 39-kb IE showed poor homology to other phosphoethanolamine transferases

The phosphoethanolamine transferases (PEA) are a family of proteins that mediate lipid A modification, and some PEA proteins encoded by mcr genes confer high levels of colistin resistance in E. coli [61]. Besides the presence of a

| Genes | Proteins/enzymes | Functions |
|-------|------------------|-----------|
| peaX² | Phosphoethanolamine transferase (PEAX)³ | LPS modification/colistin resistance |
| dndBCDE | DNA sulfur modification proteins⁴ | DNA phosphorothioation (replacing non-bridging phosphate oxygen with sulfur in the DNA backbone/Oxidative stress response |
| dptFGH | DNA phosphorothioation-dependent restriction enzymes⁵ | DNA restriction of phosphorothioated DNA/foreign DNA recognition and restriction |
| proQ | ProQ | Osmoprotectant transport activator |
| Other genes | Adhesins, DNA-binding proteins, transposases, hypothetical proteins | Possible gene expression regulation, DNA recombination, and unknown functions |

²PEAX represents phosphoethanolamine transferase encoded by the peaX gene located on the 39-kb IE of NADC 6564
³Activities of these enzymes were verified by specific in vitro assays as described in Methods and Results
phosphoethanolamine transferase (referred to as PEAX encoded by the sequence 4,825,365 bp – 4,826,864 bp) on the 39-kb IE, homology analysis revealed that NADC 6564 contains four additional phosphoethanolamine transferases (PEA) encoded at different regions (PEA1: 2,157,481 bp – 2,159,100 bp; PEA2: 2,177,878 bp – 2,19,569 bp; PEA3: 3,672,079 bp – 3,673,662 bp; and PEA4: 4,985,684 bp – 4987,327 bp) of its chromosome. The reference strains EDL933 and Sakai carry three and four genes encoding for PEA in their chromosomes, respectively. The amino acid sequence homology comparison revealed that among the five phosphoethanolamine transferases (PEA1, PEA2, PEA3, PEA4 and PEAX) of strain NADC 6564, PEA1 and PEA4 shared higher homology between their amino acid sequences (46% identity). The amino acid sequence of PEA1 also showed higher homology to the amino acid sequences of PEA transferase encoded by mcr3 (58%) and mcr1 and mcr2 (45%) (Fig. 4d). On the other hand, amino acid sequences of PEA3, PEA4, and PEAX showed lower homology to each other (32–34% identity) and to the amino acid sequences of PEA encoded by the three mcr genes (Fig. 4d). Since one of the phosphoethanolamine transferases (PEAX) is encoded by the 39-kb IE, we compared the nucleotide sequence of the peaX ORF (1499 bp) against the GenBank database to determine its homology to the nucleotide sequences of pea ORFs of E. coli strains and other bacterial species. The results of this analysis showed that peaX ORF was highly homologous to pea ORFs of E. coli strains PCN033 and PCN061 (100%) [77], avian pathogenic E. coli strain O1 (99%) (GenBank Accession Number: CP028310.1), uropathogenic E. coli strain UT189 (99%) (GenBank Accession Number: CP000243.1), and enterohemorrhagic E. coli strain 104:H21 (99%) [80]. The minimum inhibitory concentration (MIC) of colistin for NADC 6564 containing five phosphoethanolamine...
transfersases was slightly but insignificantly ($p > 0.05$) higher (0.66 μg/mL) compared to the colistin MIC (0.41 μg/mL) in reference strains (EDL933, Sakai, and a non-pathogenic E. coli TOP10) carrying three to four phosphoethanolamine transfersases (Fig. 4e). Whether the small increase in colistin resistance of NADC 6564 was due to the presence of the additional phosphoethanolamine transferase (PEAX) encoded by the 39-kb IE was not determined because overall NADC 6564 still remained colistin sensitive like the reference strains. A phylogenetic tree constructed by comparing the homology of genes (pea1, pea2, pea3, pea4, and peaX) encoding five phosphoethanolamine transfersase of NADC 6564 to genes encoding phosphoethanolamine transferases in other bacterial species revealed that all five pea genes shared a common ancestor but diverged independently into separate clades over time (Additional file 5: Figure S3).

**Comparative genomics facilitated inferring the evolutionary relationship of NADC 6564**

In order to infer the evolutionary relationship of NADC 6564 to other O157 and non-O157 strains, we constructed two types of maximum-likelihood trees using IQ-TREE and different sets of data as input. To generate the first tree, we selected 2000 genes from 40 different pathogenic and non-pathogenic E. coli strains to construct a concatenated core genome representative of these strains. The evolutionary tree generated by using the concatenated core genome (Fig. 5a) grouped O157 strains into four clades that are colored blue (clade 1), green (clade 2), purple (clade 3), and black (clade 4). The strain NADC 6564 (shown in red font) grouped with clade 1 O157 strains. This grouping agreed with the previously reported evolutionary groupings of O157 strains using octamer-based genome scanning [21]. To further confirm the validity of the tree shown in Fig. 5a, we also constructed an evolutionary tree by comparing single nucleotide polymorphism (SNP) profiles generated for the whole genome sequences of O157 and non-O157 strains using annotated genome of NADC 6564 as a reference. The tree generated from SNP profiles comparison assigned O157 strains into four groups (Fig. 5b), and this grouping was similar to that generated for O157 strains using a concatenated core genome as a reference (Fig. 5a). However, the evolutionary tree shown in Fig. 5b suggested that NADC 6564 might have diverged much earlier from other lineage I O157 strains included in this tree.

**Discussion**

*Escherichia coli* O157:H7 (O157) has become a prominent foodborne pathogen not only in North America but worldwide since the first reported association O157 with human disease in early 1980s [12, 81]. O157 has been called an emerging foodborne pathogen due to the continuous evolution of its genome via acquisition of laterally transferred mobile DNA elements and other mutagenic events [26, 30, 45, 72]. Many of the newly emerged clones of O157 that have caused outbreaks inflicting high morbidity and mortality in infected individuals have in part been linked to increased virulence of the implicated O157 strains [27]. Thus, understanding genomic diversity and plasticity of O157 strains is important for determining epidemiology, understanding bacterial pathogenesis, identifying specific biomarkers, predicting severity of the disease, tracing origin of these strains, and developing vaccines for controlling these pathogens in host reservoirs.

In a previous study, we reported basic features of the genome of a foodborne isolate of O157 strain 86–24 linked to the 1986 human disease outbreak [28]. In the present study, we used comparative genomics to analyze chromosomal sequence of this isolate, named NADC 6564, to determine its genetic content and organization, unique functional attributes, genetic lineage, and evolutionary relationship to other well-characterized O157 strains. The genetic contents, in terms of the total nucleotides constituting the chromosome of NADC 6564 (5.46 Mb), was slightly smaller than the reference strain O157 EDL933 (5.53 Mb) and O157 Sakai (5.49 Mb). Although all *E. coli* share about 4.1 Mb core genome sequence, in pathogenic *E. coli* strains, such as Shiga toxin-producing *E. coli* (STEC) O157:H7 and non-O157 STEC strains, the core genome is disrupted by the insertion of laterally-acquired mobile DNA elements. The acquisition of variable numbers and types of insertion elements, such as bacteriophages (phages), transposons, and genomic islands (GI), is one of the major factors in genome size variability among different O157 and non-O157 strains [47, 72, 82]. For example, screening of the whole chromosomal sequence for inserted phages and GI revealed the presence of 19 phages (named P1 – P19) and 53 GI in NADC 6564 compared to the presence of 18 phages and 63 GI in reference strain EDL933 and 18 phages and 71 GI in the reference strain Sakai. These laterally-acquired elements would account for about 1.645 Mb of DNA in NADC 6564 compared to 1.641 Mb and 1.372 Mb of acquired DNA in the two reference strains EDL933 and Sakai, meaning that these strains would have 3.815 Mb, 3.889 Mb, and 4.118 Mb, respectively of core *E. coli* DNA as part of their genomes. The reported core *E. coli* DNA for the two reference O157 strains is about 4.1 Mb [83].

A variety of high resolution molecular subtyping methods have been developed for discriminating O157 strains and inferring their evolutionary relationships. Some of these methods have classified O157 strains into three lineages, human-biased lineage I strains, bovine-biased lineage II strains, and lineage I/II strains, which are ambiguous with
respect to host association and phenotype [21, 45, 84]. We used available bioinformatics tools to construct evolutionary trees based on concatenated core genome and SNPs of NADC 6564 and the reference O157 strains. In these two trees, str. NADC 6564 and the reference O157 strains were grouped correctly, which was in agreement with previous studies [21, 45, 84], into lineage I, II, and I/II strains.

However, in the tree constructed by comparing concatenated core genome, NADC 6564 appeared to be closely related to lineage I strain EDL933 as these two shared a common ancestral strain. In the tree generated by comparing SNP profiles that took into consideration 58,505 SNPs, NADC 6564 is more distantly related to the lineage I reference strains (EDL933, Sakai, XuZhou21, WS4202, and

---

**Fig. 5** Cladograms showing evolutionary relationship of E. coli O157 strain NADC 6564 to other pathogenic O157 and non-O157 strains. 

a) Cladogram produced by constructing concatenated core genomes from whole genome sequences of 40 bacterial strains listed in Additional file 6: Table S1. A total of 2284 core genes were identified among the 40 strains used for constructing core genomes. These core genomes were used to infer a maximum likelihood phylogenetic tree using IQTREE. The generated tree was visualized using FigTree. Since there was no outgroup included in the tree, FigTree’s midpoint rooting method was used to root the tree.

b) Cladogram inferred using the parSNP-Gingr pipeline. Whole genomes for the sequences being compared were provided as input to parSNP using the NADC 6564 annotated GenBank file as the reference. Evolutionary trees were created by visualizing the parSNP output in Gingr.
The identification of a unique 39-kb IE, altered chromosomal organization, and a distinctive phage profile provide important supporting evidence that the NADC 6564 might also have a distinctive SNP profile compared to the other lineage I strains, and warrant characterization of these SNP differences in future studies.

Besides observing subtle differences in the total amount of acquired DNA between NADC 6564 and reference strains EDL933 and Sakai, the organization of the core chromosomal sequence of strain NADC 6564 differed strikingly from the reference strains. We found that 1.4 Mb of the genome represented by contiguous 18 linear conserved blocks or LCB were inverted in their orientation in the chromosome of strain NADC 6564 compared to reference O157 and non-O157 strains. This disparate organization of the large portion of the chromosome would suggest occurrence of specific recombination events presumably resulting in the inversion of this large section of the chromosomal sequence in NADC 6564. Since the laterally-acquired DNA elements in O157 strains are or were mobile elements, the differential integration and subsequent recombination events, such as truncations and excisions, in the chromosome could account for observed variations in the arrangement of conserved core genome blocks among O157 strains [72, 82, 85].

The laterally-acquired elements most often carry genes that could impact the virulence attributes of bacterial strains, and this phenomenon is well exemplified in emergence and evolution of O157 and non-O157 STEC strains. It has been proposed that STEC O157:H7 strains evolved from an ancestral strain similar to enteropathogenic E. coli (EPEC) O55:H7 that had already acquired a locus of enterocyte effacement (LEE) for adherence to and colonization of human intestinal tract [52, 53]. This LEE-carrying ancestral strain in the distant past diverged into EPEC O55:H7 and STEC O157:H7 lineages. STEC O157:H7 lineage resulted from lateral acquisition of Shiga toxin-encoding bacteriophages, new genomic islands (GI) encoding O157-specific O antigens, and several other GI and bacteriophages impacting virulence and metabolic repertoire of STEC O157:H7 isolates [52, 53, 72, 82]. Whole genome sequence analysis of NADC 6564 confirmed the presence of both the Shiga toxin 2-encoding bacteriophage (numbered P6; Fig. 2) and LEE-encoded pathogenicity island inserted in the vicinity of GI 15 and GI 3, respectively. The Stx2 phage and LEE were inserted in the chromosome of NADC 6564 adjacent to wbrA and tRNA-Sec, respectively. These two elements are also inserted at the same sites in EDL933 and Sakai strains [86, 87]. In addition, we identified a 39-kb DNA region bearing features of an insertion element (IE), such as the presence of a 56-bp direct repeat at its ends, genes for integrase and transposases, a phage remnant (labeled P19), and use of tRNA-leu integration site in the NADC 6564 chromosome. This 39-kb IE is inserted in LCB 31, and based on BLAST homology search none of the O157, non-O157, and nonpathogenic E. coli K12 used as reference strains harbored this IE in their genomes (Fig. 2). However, genomic regions ranging in size from 30 to 37 kb, showing 100% homology to the corresponding regions of the 39-kb IE of NADC 6564, were identified in a porcine E. coli strain PCN061 [77] and unpublished genomes of E. coli strain AR0015 (Accession Number: CP024862.1) and E. coli strain C1 (Accession Number: CP010116.1). Most other E. coli strains showed only a limited homology (50% or less) to the 39-kb IE of NADC 6564. These results suggest that a few E. coli strains including NADC 6564 had acquired this IE through lateral gene transfer while the other E. coli strains either had not acquired or had this IE deleted during the course of evolution.

Besides carrying genes for integrase and transposases, the 39-kb IE of NADC 6564 carried a 10.1 kb phage remnant and genes encoding for an adhesin, phosphoethanolamine transferase, S-DNA modification and a DNA restriction system that could impact virulence and survival of NADC 6564. For example, the homologs of the adhesin gene present on the 39-kb IE of NADC 6564 were identified in E. coli PCN061 of porcine origin [77], uropathogenic E. coli strain UT189 (Accession number CP000243.1), and enteroaggregative E. coli strain O104:H21 that was linked to a major disease outbreak in Europe in 2011 affecting 4321 people, of which 852 developed hemorrhagic uremic syndrome and 53 deaths [88].

The phosphoethanolamine transferases are a family of proteins that mediate outer membrane lipid A modification and some members of this family that are encoded by the mcr genes confer high levels of colistin resistance in E. coli [61]. It has been suggested that the modification of lipid A by the chromosomally encoded phosphoethanolamine transferases could result in a slight increase in the minimum inhibitory concentration (MIC) of colistin and other cationic antibiotics that could confer survival advantage to E. coli O157:H7 in certain environmental niches [62]. We were able to demonstrate that MIC of colistin in NADC 6564 containing genes (pea1, pea2, pea3, pea4, and peaX) encoding five phosphoethanolamine transferases (PEA1 – PEA4 and PEAX), was slightly elevated compared to the reference strains containing genes encoding only three or four phosphoethanolamine transferases. Based on evolutionary tree analysis of pea and peaX genes (Additional file 5: Figure S3), the peaX gene (encoded by the 39-kb IE) of NADC 6564 might have evolved in an E.
coli-type ancestral strain and subsequently acquired by O157 strains, such as NADC 6564, very early in their evolution compared to the acquisition of other pea genes by these strains. Evolutionary tree data and the observed low homology of peaX gene to other pea genes suggest that other pea genes have undergone greater divergence in their nucleotide sequences during the course of evolution compared to the peaX gene since their divergence from the first common ancestral strain.

The dndBCDE and dptFGH operons located on the 39-kb IE in NADC 6564 were functionally active as both S-DNA modification and restriction of DNA lacking S-DNA modification were observed. The dndBCDE and dptFGH operons are widely distributed in diverse bacterial species [57], but we for the first time have identified these operons in an O157 strain. In addition, demonstration of the presence of dnd genes in genomic islands in a majority of dnd-carrying E. coli strains, including NADC 6564 as we have described in the current study, indicates the importance of these islands in horizontal transfer of dnd genes [58]. Many bacterial species possess both dndBCDE and dptFGH operons organized in a same order but some bacterial species only harbor the dndBCDE genes [57, 89]. Since the dndBCDE operon has been shown to confer protection to bacterial hosts from a variety of stressors, especially oxidative and high temperature stressors through S-modification of DNA [56], it is feasible that NADC 6564 carrying these genes might have better survival advantages under conditions of high oxidative stress that could be induced in the gastrointestinal tract of the host reservoir animal or in the gastrointestinal tract of infected humans by a variety of factors. Evolutionary tree analysis of dndBCDE and dptFGH operons present in diverse bacterial taxa appeared to suggest that NADC 6564, like many other E. coli strains that it clusters with [58], might have acquired these genes very early in the evolution and further dissemination of dndBCDE and dptFGH genes occurred later to other bacterial species (Additional file 3: Figure S1; Additional file 4: Figure S2).

Conclusion
In summary, we have utilized comparative genomics to provide insights into the genetic and functional organization of the whole chromosomal sequence of a foodborne isolate of E. coli O157:H7 strain NADC 6564. By doing so, we discovered not only a unique phage distribution profile and a novel insertion element encoding a set of functions relevant to enhanced stress tolerance, but also identified some disparities in its evolutionary relationships with other closely related lineage I strains. Characterization of the SNP profile and SNP effects on specific phenotypes of NADC 6564 would provide a greater understanding of the microevolutionary history behind the observed evolutionary relationship of NADC 6564 to other lineage I strains. In addition, studying the effects of SNPs on gene expression and determining whether the acquired stress related functions confer host and/or environmental fitness would help predict virulence potential and ability to survive in growth-limiting environments.

Methods
Construction of circular chromosomal maps
The chromosomal maps of bacterial strains were created and compared to each other utilizing the BLAST Ring Image Generator (BRIG). BRIG uses BLAST to map homologous regions from the query sequences to a circular map of the reference sequences [75, 90, 91]. The annotated chromosome of E. coli O157:H7 NADC 6564 [28] was used as a reference for generating whole chromosomal sequence comparisons with selected query sequences. BRIG was also used to generate detailed maps of the regions that were uniquely identified in NADC 6564 but were absent in query sequences. The names and accession numbers of sequences used in this study were downloaded from the NCBI database (Additional file 6: Table S1).

Identification of genomic islands, phage or phage-like elements, and DNA repeats
The location of putative phage or phage-like regions in the chromosome of NADC 6564 were determined by downloading a FASTA file of the whole chromosomal sequence of this strain from GenBank and then uploading this file to PHASTER [73, 74]. To determine if phage regions identified in NADC 6564 have homologous counterparts in reference strains EDL933 and Sakai, we BLAST searched NADC 6564 phage regions against the counterparts in reference strains EDL933 and Sakai, we BLAST searched NADC 6564 phage regions against the chromosomal sequences of these two reference strains. Genomic islands (GI) in NADC 6564 were predicted using the online IslandViewer4 [92]. IslandViewer4 integrates the IslandPath-DIMOB, SIGI-HMM, and IslandPick prediction methods to find probable locations of GI. Since multiple methods are integrated by IslandViewer4 when determining start and end positions of GI, sometimes these GI predictions overlapped. In this situation, the overlapping GI were combined. The GI labeled on the BRIG map were merged in this fashion. The nucleotide sequence of each genomic island of NADC 6564 was also used as a query against the published chromosomal sequences of reference strains EDL933 and Sakai to determine the presence and location of homologous GI in the reference strains. Exact pairs of repeats and exact tandem repeats were discovered in the chromosome of NADC 6564 by utilizing Mummer 3.23 [93]. Repeat-match program of mummer package identifies maximal exact repeats within a single sequence. This program simply exploits Mummer’s ability to align two different sequences based on sequence homology by
applying it to one single sequence. The minimum match length option was set to 50 nucleotides.

**Whole chromosomal sequence alignments**

In order to construct and visualize whole chromosomal alignments, a multiple genome alignment tool called Mauve was used [94]. Mauve compares multiple genome sequences and finds regions of homology called locally collinear blocks (LCB). The progressive Mauve algorithm was used with the default parameters in order to generate an alignment file in XMFA format. This XMFA file was then visualized using the Mauve GUI. Upon inspection of the initial Mauve alignment, the beginning of the chromosomal sequence of NADC 6564 [28] contained 6 LCB that were found at the end of the genomes in the other strains. In order to rearrange the LCB so that all of the sequences are uniform, the XMFA file was manually parsed to find the exact location of the first six LCB. The chromosomal sequence of strain 6564 was then modified so that the nucleotide sequence of the first six LCB were moved to the end.

**Evolutionary trees**

**Core genome comparison**

In order to create a meaningful phylogenetic tree comparing NADC 6564 to other *E. coli* strains, the core genomes of 40 different strains were compared. To accomplish this, a FASTA file containing a list of genes and their nucleotide sequences was downloaded from the NCBI database for each genome compared (Additional file 6: Table S1). Next, all of the FASTA files downloaded were then concatenated into one FASTA file containing a list of all of the genes and respective nucleotide sequences from all of the strains being compared. This FASTA file was used as input for the program CD-HIT in order to cluster these genes into groups of equal length with similarity of 90% and above [95, 96]. The resulting clusters were filtered so that only the clusters containing exactly one gene from each chromosomal genome remained. This ensures that the gene is a core gene present in all of the strains compared. A total of 2284 core genes were identified among the 40 strains being compared. Once the clusters containing core genes were isolated, the clusters were split so that a FASTA file could be created for each strain containing core genes from the respective strain. These files are not identical because CD-HIT isolated the genes that are 90% homologous. Although the FASTA files aren’t identical, all of the FASTA files contain the same set of genes with the nucleotide sequence allowed to vary by 10% at most when compared across genomes. These core genomes were combined into a single multiFASTA file and used to infer a maximum likelihood phylogenetic tree using IQ-TREE [97]. The IQ-TREE phylogenetic inference consisted of using the model finder option (–m TESTNEW) to find the best evolutionary model appropriate for the input data [98]. Once the GTR + R4 was identified as the best evolutionary model, a maximum likelihood tree was constructed using IQ-TREE. Support values drawn by IQ-TREE were found using 1000 bootstrap replicates in combination with both the ultrafast bootstrap approximation method as well as the SH-like approximate ratio test method [99, 100]. The generated tree was visualized using FigTree using the midpoint rooting method [101]. Since there was no outgroup included in the tree, FigTree’s midpoint rooting method was used to root the tree.

**ParSNP-Gingr**

Core phylogeny was also inferred using the parSNP-Gingr pipeline [102]. Whole genome sequences being compared were provided as input to parSNP using the NADC 6564 annotated GenBank file as the reference. Trees were created by visualizing the parSNP output in Gingr [102].

**Determination of DNA backbone S-modification (phosphorothioation) activity**

The DNA backbone S-modification was assessed by treating genomic DNA of strains carrying (NADC 6564 and NADC 6565) or lacking (EDL933 and Sakai) dndBCDE operon, which encode enzymes for this activity [57], with 30 mM iodine solution for 15 min at 60 °C. The treated DNA was cooled to 4 °C and analyzed on a 1% agarose gel by a standard agarose gel electrophoresis procedure [57]. The strain NADC 6565 is a Congo red-positive variant of NADC 6564 and it also carries the dndBCDE operon based on the published genome sequence of these two strains [28]. Therefore, strain NADC 6565 was used as a positive control for confirming the presence of S-DNA modification activity in other strains tested in this assay.

**Determination of dpt encoded restriction system activity against DNA lacking S-modification**

A high copy plasmid (pUC19) lacking the S-DNA modification was transformed by electroporation into *E. coli* strains harboring or lacking both dndBCDE and dptFGH operons, which encode DNA S-modification and a restriction system for restricting DNA lacking the S-modification, respectively. Electroporations were performed using a Gene Pulser (Bio-Rad Laboratories, Inc., Hercules, CA) and according to the manufacturer’s instructions. The transformed *E. coli* cells were diluted 1:10 in SOC (Super Optimal broth with Catabolite repression) broth (Bio 101, Inc., La Jolla, CA) and incubated for one hour at 37 °C on a shaker set to 200 rpm. Ten-fold serial dilutions of these cell suspensions were spread-plated on LB agar containing carbenicillin (100 μg/mL) and plates were incubated overnight at 37 °C. The number of colonies produced on the
overnight-incubated plates from three independent experiments were counted and plotted.

Determination of minimum inhibitory concentration (MIC) of colistin
Bacterial strains were grown overnight in LB broth (strains Sakai, EDL33, and TOP10), LB broth containing 100 μg/mL streptomycin (strain NADC 6564), or LB broth containing 24 μg/mL of colistin (E. coli strain BEAR 119605; kindly provided by Dr. Heather Allen, Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, ARS-USDA, Ames, Iowa, USA) at 37 °C on a shaker (200 rpm). The overnight cultures were diluted to 5 × 10^6 cells/mL in Mueller Hinton Broth (MHB) containing 12.5 mg/L magnesium chloride and 12.5 mg/L calcium chloride (MHB-MC). Aliquots (100 μl) of diluted culture were added to the wells of a 96-well plate containing 100 μL of MHB-MC or MHB-MC supplemented with colistin (starting at 0.1 μg/mL in the first well and increasing by an order of 2-fold in succeeding wells so that the last well of the row contained 64 μg/mL of colistin). The plates were incubated at 37 °C for 18 to 24 h and wells were examined for any visible bacterial growth. The wells with the lowest concentration showing no visible bacterial growth were accepted as the MIC of colistin for that particular bacterial strain.

Determination of evolutionary relationship of genes encoding phosphoethanolamine transferases and proteins mediating S-DNA modification (Dnd proteins) and restriction of unmodified DNA (Dpt proteins)
The phylogenetic trees displaying the evolutionary similarity of phosphoethanolamine transferases, Dnd, and Dpt proteins encoded in NADC 6564 were generated using IQ-TREE [97]. To construct an evolutionary tree of phosphoethanolamine transferases, the amino acid sequences for respective proteins were aligned using the MUSCLE alignment tool with default parameters [104]. The alignment files for each respective protein search was used as input for IQ-TREE [97]. IQ-TREE phylogenetic inference consisted of using the -m TEST-NEW option in order to identify the best-fit evolutionary model according to the Bayesian Information Criterion [98]. Once GTR + I + G4 was identified as the best evolutionary model for both Dpt and Dnd data, a maximum likelihood tree was generated using IQ-TREE. Support values drawn by IQ-TREE were found using 1000 bootstrap replicates in combination with both the ultrafast bootstrap approximation method as well as the SH-like approximate ratio test method [99, 100]. All phylogenetic trees were visualized in FigTree using the midpoint rooting method if no outgroup was provided [101].

Statistical analyses
A two sample, Student’s t-test was used to determine the significance of the difference in the number of carbencillin-resistant colonies recovered from NADC 6564 and EDL933 after transformation with pUC19. Statistical significance of the difference in the minimum inhibitory concentration of colistin between NADC 6564 and the reference strains was assessed using one-way analysis of variance with multiple comparison of means. Data were analyzed with GraphPad Prism7 software (GraphPad Software, La Jolla, CA). The difference was considered significant at p < 0.05.

Disclaimer
Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Additional files

Additional file 1: Table S2. Sequence length and location of chromosomal regions in two reference strains exhibiting homology to bacteriophage regions of NADC 6564. (DOCX 33 kb)
Additional file 2: Table S3. Chromosomal location of genomic islands (GI) in NADC 6564 and corresponding homologs in two reference strains. (DOCX 38 kb)
Additional file 3: Figure S1. Cladograms showing evolutionary relationship of the dndBCDE gene sequences in NADC 6564 to other bacterial species. The nucleotide sequences of dndBCDE genes was used as a query (indicated in the red font) to download homologs of these genes. These homologous sequences were then used for constructing a maximum likelihood phylogenetic tree using IQ-TREE. The generated tree was visualized using FigTree. (DOCX 434 kb)
Additional file 4: Figure S2. Cladograms showing evolutionary relationship of the dptFGH gene sequences in NADC 6564 to other bacterial species. The nucleotide sequences of dptFGH genes was used as a query (indicated in the red font) to download homologs of these genes. These homologous sequences were then used for inferring a maximum likelihood phylogenetic tree using IQ-TREE. The generated tree was visualized using FigTree. (DOCX 293 kb)
Additional file 5: Figure S3. Cladograms showing evolutionary relationship of the peaX gene sequences of NADC 6564 to other bacterial species. The nucleotide sequences of peaX genes was used as a query (indicated in the red font) to download homologs of these genes. These homologous sequences were then used for constructing a maximum likelihood phylogenetic tree using IQ-TREE. The generated tree was visualized using FigTree. (DOCX 234 kb)
Additional file 6: Table S1. Bacterial strains with their corresponding accession numbers used in the study. (DOCX 29 kb)
Acknowledgements
We thank Lindsay Andersen for technical support in the completion of this study, Dr. Crystal Loving for a critical review of the manuscript, Dr. Nicole Ricker and Daniel Nielsen for helpful discussions on mor family of genes and evolutionary tree analysis, and Dr. Heather Allen for supplying colistin-resistant E. coli strain. This work was supported by USDA-ARS CRIS project 5030-32000-112-00D. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Funding
This work was supported by USDA, ARS CRIS funds. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Availability of data and materials
The complete chromosomal sequence of NADCC 6564 is available at the GenBank under the assigned accession number CP017251. Other chromosomal sequences used for comparative genomics were retrieved from GenBank using their accession numbers that are listed in Additional file 6: Table S1. E. coli O157:H7 strain NADCC 6564 will be provided pending that the requester would fulfill requirements for shipment of RG2 bacterial agents.

Authors’ contributions
VS designed the study, performed experiments, and prepared the draft of the manuscript; SA performed comparative genomic analysis, performed experiments, and contributed to manuscript writing; RS performed experiments and reviewed manuscript; DB analyzed data and reviewed the manuscript. All authors read and approved the final draft of the manuscript.

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Food Safety and Enteric Pathogens Research Unit, USDA, ARS, National Animal Disease Center, 1920 Dayton Avenue, P.O. Box 70, Ames, IA 50010, USA. 2Infectious Bacterial Diseases Research Unit, National Animal Disease Center, ARS-USDA, Ames, Iowa, USA. 3Oak Ridge Institute for Science and Education (ORISE), ARS Research Participation Program, MS 36, P.O. Box 117, Oak Ridge, TN 37831, USA.

Received: 27 September 2018 Accepted: 25 February 2019
Published online: 08 March 2019

References
1. Ackers ML, Mahon BE, Leahy E, Goode B, Damrow T, Hayes PS, Bibb WF, Rice DH, Barrett TJ, Huttuganer L, et al. An outbreak of Escherichia coli O157:H7 infections associated with leaf lettuce consumption. J Infect Dis. 1998;177(6):1588–93.
2. Bell BP, Goldoft M, Griffin PM, Davis MA, Gordon DC, Tarri J, Bartleson CA, Lewis JH, Barrett TJ, Wells JG, et al. A multistate outbreak of Escherichia coli O157:H7-associated bloody diarrhea and hemolytic uraemic syndrome from hamburgers. The Washington experience. JAMA. 1994;272(17):1349–53.
3. Hillborn ED, Mermin JH, Mshar PA, Hadler JL, Voetsch A, Wolfkunzi C, Swartz M, Mshar R, Lambert-Fair MA, Farrar JA, et al. A multistate outbreak of Escherichia coli O157:H7 infections associated with consumption of mesclun lettuce. Arch Intern Med. 1999;159(15):1758–64.
4. Keene WE, Hedberg K, Henriott DE, Hancock DD, McKay RW, Barrett TJ, Fleming DIW. A prolonged outbreak of Escherichia coli O157:H7 infections caused by commercially distributed raw milk. J Infect Dis. 1997;176(3):815–8.
5. Rangel JM, Sparrling PH, Crowe C, Griffin PM, Swerdlow DL. Epidemiology of Escherichia coli O157:H7 outbreaks, United States, 1982–2002. Emerg Infect Dis. 2005;11(6):693–9.
6. Banavali N, Griffin PM, Greene KD, Barrett TJ, Bibb WF, Green JH, Wells JG. Hemolytic uraemic syndrome study C. The United States National Prospective Hemolytic Uremic Syndrome Study: microbiologic, serologic, clinical, and epidemiologic findings. J Infect Dis. 2001;183(7):1063–70.
7. Segler R, Oakes R. Hemolytic uraemic syndrome: pathogenesis, treatment, and outcome. Curr Opin Pediatr. 2005;17(2):200–4.
8. Hancock D, Besier T, LeJeune J, Davis M, Rice DH, Osterman MEK. The control of VTEC in the animal reservoir. Int J Food Microbiol. 2001;66(1):271–81.
9. CDC. Escherichia coli O157:H7 infection in children associated with raw milk and raw colostrum from cows—California. MMWR. 2006;55(23):625–8.
10. Elder RO, Keen JE, Srigursa GR, Barkocy-Gallagher GA, Kooshmarae M, Laegreid W. Correlation of enterohemorrhagic Escherichia coli O157 prevalence in feces, hides, and carcasses of beef cattle during processing. Proc Natl Acad Sci U S A. 2000;97(7):3999–3003.
11. Gyles CL. Shiga toxin-producing Escherichia coli: an overview. J Anim Sci. 2007;85(13 Suppl):E45–62.
12. Riley LW, Remis RS, Helgerson SD, Gigee HB, Wells JG, Davis BR, Hebert RJ, Ollott ES, Johnson LM, Haigrett NT, et al. Hemorrhagic colitis associated with a rare Escherichia coli serotype. N Engl J Med. 1983;308(12):681–5.
13. System C-NOR. Disease outbreak data for 1998–2016. 2018.
14. Davis MA, Hancock DD, Besier TE, Call DR. Evaluation of pulsed-field gel electrophoresis as a tool for determining the degree of genetic relatedness between strains of Escherichia coli O157:H7. J Clin Microbiol. 2003;41(5):1843–9.
15. Davis MA, Hancock DD, Besier TE, Rice DH, Hovde NJ, Guldason RO, Duffield AK, Call DR. Correlation between geographic distance and genetic similarity in an international collection of bovine faecal Escherichia coli O157:H7 isolates. Epidemiol Infect. 2003;131(2):523–30.
16. Elhadidy MM, Elkhatabi WF. Multilocus genotypic characterization of Escherichia coli O157:H7 recovered from food sources. Epidemiol Infect. 2015;143(11):2367–72.
17. Noller AC, McEllistrem MC, Harrison LH. Genotyping primers for fully automted multilocus variable-number tandem repeat analysis of Escherichia coli O157:H7. J Clin Microbiol. 2004;42(8):3907.
18. Mellor GE, Besier TE, Davis MA, Beavis S, Jung W, Smith HV, Jennison AV, Doyle CJ, Chardy PS, Gobius KS, et al. Multilocus genotype analysis of Escherichia coli O157 isolates from Australia and the United States provides evidence of geographic divergence. Appl Environ Microbiol. 2013;79(16):5050–8.
19. Bustamante AV, Lucchesi PM, Parra AE. Molecular characterization of Verocytotoxigenic Escherichia coli O157:H7 isolates from Argentina by multiple-loci VNTR analysis (MLVA). Braz J Microbiol. 2009;40(4):927–32.
20. Heuvink AE, van de Kar NC, Meis JF, Monnens LA, Melchers WJ. Characterization of verocytotoxin-producing Escherichia coli O157 isolates from patients with haemolytic uraemic syndrome in Western Europe. Epidemiol Infect. 1995;115(1):1–14.
21. Kim J, Niewfjeld J, Benson AK. Octamer-based genome scanning distinguishes a unique subpopulation of Escherichia coli O157:H7 strains in cattle. Proc Natl Acad Sci U S A. 1999;96(23):13288–93.
22. Ziebell K, Steele M, Zhang Y, Benson A, Taboada EN, Laing C, McEwen S, Dyer DJ, Chandy PS, Gobius KS, et al. Multilocus genotype analysis of Escherichia coli O157:H7 isolates from Argentina by whole genome sequencing. Microbial Genom. 2016;2(10):8.
23. Zhang W, Qi W, Albert TJ, Motiwala AS, Alland D, Hyytia-Trees EK, Ribot EM, Fields PI, Whittam TS, Swanmanathan B. Probing genomic diversity and evolution of Escherichia coli O157 by single nucleotide polymorphisms. Genome Res. 2006;16(6):757–67.
67. Sharma VK, Bearson BL. Hnf controls Escherichia coli O157:H7 biofilm formation by differential regulation of global transcriptional regulators FihDC and CsgD. Appl Environ Microbiol. 2013;79(7):2384–96.
68. Sharma VK, Casey TA. Escherichia coli O157:H7 lacking the qseBC-encoded quorum-sensing system outcompetes the parental strain in colonization of cattle intestines. Appl Environ Microbiol. 2014;80(6):1882–92.
69. Sharma VK, Kudva IT, Bearson BL, Stasko JA. Contributions of EspA filaments and Curl fimbriae in cellular adherence and biofilm formation of Enterohemorrhagic Escherichia coli O157:H7. PLoS One. 2016;11(2):e0149745.
70. Sheng H, Nguyen YN, Hovde CJ, Sperandio V. SdiA aids enterohemorrhagic Escherichia coli carriage by cattle fed a forage or grain diet. Infect Immun. 2013;81(9):2947–8.
71. Sinclair JF, Dean-Nystrom EA, O’Brien AD. The established intimin receptor Tir localize at or near the site of enterohemorrhagic Escherichia coli O157:H7 adherence to enterocytes in vivo. Infect Immun. 2006;74(2):1255–65.
72. Ogura Y, Ooka T, Iuchi A, Toh H, Aida-Murakami M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, et al. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci U S A. 2009;106(4):17393–44.
73. Amr D, Grant JR, Marcus A, Sajed T, Pon A, Liang Y, Wishart DS. PHAST: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44(W1):W16–21.
74. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39(Web Server issue):W547–52.
75. Alikhani NF, Petry NK, Ben Zakour NL, Beatson SA. BLAST ring image generator (BRIG): simple prokaryote-genome comparisons. BMC Genomics. 2011;12:402.
76. Carter MQ, Louie JW, Huynh S, Parker CT. Natural isps mutations contribute to population heterogeneity in Escherichia coli O157:H7 strains linked to the 2006 US spinach-associated outbreak. Food Microbiol. 2014;44:108–18.
77. Liu C, Zheng H, Yang M, Xu Z, Wang X, Wei L, Tang B, Liu F, Zhang Y, Ding Y, et al. Genome analysis and in vivo virulence of porcine extraintestinal pathogenic Escherichia coli strain PCN333. BMC Genomics. 2015;16:717.
78. An X, Xiong W, Yang Y, Li F, Zhou X, Wang Z, Deng Z, Liang J. A novel target of IscS in Escherichia coli participating in DNA phosphorothiation. PLoS One. 2012;7(12):e51265.
79. Ou HY, He X, Shao Y, Tai C, Rajakumar K, Deng Z, dxdDB. A database focused on phosphorothiation of the DNA backbone. PLoS One. 2009;4(6):e51262.
80. Gonzalez-Escalona N, MA MF, Rump LV, Payne J, Andrzejewski D, Brown EW, Ohtsubo E, Nakayama K, et al. Complete genome sequence of Escherichia coli K-12 MG 1655 defines the genetic requirements for STEC virulence. J Bacteriol. 2008;190(3):955–65.
81. Majowicz SE, Scallan E, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Copeland KL, Baggs S, May ML, et al. Isolates Causing Hemorrhagic Colitis during a 1994 Montana outbreak in vivo. Infect Immun. 2006;74(1):329–36.
82. Majowicz SE, Scallan E, Shun A, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Copeland KL, Baggs S, May ML, et al. Comparison of STEC phylogenies: are bacteriophages the major contributors? Trends Microbiol. 2009;17(1):22–8.
83. Majowicz SE, Scallan E, Shun A, Jones-Bitton A, Sargeant JM, Stapleton J, Angulo FJ, Copeland KL, Baggs S, May ML, et al. Comparison of STEC phylogenies: are bacteriophages the major contributors? Trends Microbiol. 2009;17(1):22–8.
84. rails AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
85. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
86. Kurtz S, Phillippy A, Delcher AL, Dew I, Shu S, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
87. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
88. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
89. Kurtz S, Phillippy A, Delcher AL, Dew I, Shu S, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
90. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
91. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
92. Bertelli C, Laird MR, Williams KP. Simon Fraser University research computing G, Lau BY, Hoad G, Winsor GL, brinkman FSL. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. 2017;45(W1):W30–5.
93. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway K, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
94. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
95. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658–9.
96. Kurtz S, Phillippy A, Delcher AL, Dew I, Shu S, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
97. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
98. Kalyaanamoorthy S, Minh BQ, Wink TFK, van Hae塞尔er A, Jerlins LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.
99. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.
100. Haas BT, Geiser DM, Geiser JD, de Bruijn M, et al. Improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018;35(2):518–22.
101. FigTree: http://tree.bio.ed.ac.uk/software/figtree/
102. Tennyson JI, Onodvo BD, Koren S, Phillippy AM. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):524.
103. Yen W, Li H, Shen Y, Liu Z, Wang S, Shen Z, Zhang R, Walsh TR, Shen J, Wang Y. Novel Plasmid-Mediated Colistin Resistance Gene mcr-3 in Escherichia coli MB. MBio. 2017;8(3):e00543–17.
104. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1797–2.