Comparative genomics of two super-shedder isolates of *Escherichia coli* O157:H7

Robab Katani[1,2], Rebecca Cote[2,3], Indira T. Kudva[4], Chitrita DebRoy[3,5], Terrance M. Arthur[6], Vivek Kapur[1,2]*

1 Department of Animal Science, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** Department of Veterinary and Biomedical Sciences, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **4** Food Safety and Enteric Pathogens Research Unit, National Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture, Ames, Iowa, United States of America, **5** *E. coli* Reference Center, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **6** Roman L. Hruska U.S. Meat Animal Research Center, Agricultural Research Service, U.S. Department of Agriculture, Clay Center, Nebraska, United States of America

* vkapur@psu.edu

Abstract

Shiga toxin-producing *Escherichia coli* O157:H7 (O157) are zoonotic foodborne pathogens and of major public health concern that cause considerable intestinal and extra-intestinal illnesses in humans. O157 colonize the recto-anal junction (RAJ) of asymptomatic cattle who shed the bacterium into the environment through fecal matter. A small subset of cattle, termed super-shedders (SS), excrete O157 at a rate ($\geq 10^4$ CFU/g of feces) that is several orders of magnitude greater than other colonized cattle and play a major role in the prevalence and transmission of O157. To better understand microbial factors contributing to super-shedding we have recently sequenced two SS isolates, SS17 (GenBank accession no. CP008805) and SS52 (GenBank accession no. CP010304) and shown that SS isolates display a distinctive strongly adherent phenotype on bovine rectal squamous epithelial cells. Here we present a detailed comparative genomics analysis of SS17 and SS52 with other previously characterized O157 strains (EC4115, EDL933, Sakai, TW14359). The results highlight specific polymorphisms and genomic features shared amongst SS strains, and reveal several SNPs that are shared amongst SS isolates, including in genes involved in motility, adherence, and metabolism. Finally, our analyses reveal distinctive patterns of distribution of phage-associated genes amongst the two SS and other isolates. Together, the results of our comparative genomics studies suggest that while SS17 and SS52 share genomic features with other lineage I/II isolates, they likely have distinct recent evolutionary histories. Future comparative and functional genomic studies are needed to decipher the precise molecular basis for super shedding in O157.

Introduction

Shiga toxin-producing *Escherichia coli* O157:H7 (O157) is a major foodborne pathogen that causes severe illnesses in humans worldwide. Symptoms vary from bloody diarrhea with abdominal discomfort to hemorrhagic colitis, and may also include severe life-threatening complications such as hemolytic uremic syndrome (HUS) [1]. Identified as a human pathogen in 1982, the Centers for Disease Control and Prevention (CDC) has estimated that O157 infections cause 73,000 illnesses, 2,200 hospitalizations and 60 deaths annually in the United States, as well as several outbreaks and sporadic infections globally [2].

Asymptomatic cattle are the principal animal reservoir and carry the bacteria predominantly in the terminal recto-anal junction (RAJ) of the gastrointestinal tract [3, 4]. O157 adherence and colonization of the RAJ results in cattle shedding the pathogen in their feces, leading to environmental contamination and transmission of the organism, and ultimately contamination of the food supply and outbreaks of disease in humans. Typically, cattle range from being non-shedders of O157 to shedding 100 CFU/g of feces; however, subgroups of cattle, termed as "super-shedders" (SS) have been found to excrete O157 at levels of $\geq 10^4$ CFU/g of feces [3, 4]. Several epidemiological modeling studies have suggested that although the number of SS animals on farm is often less than 10%, these animals are responsible for up to 99% of the bacteria shed into the environment [5]. The shedding of this pathogen at several magnitudes higher than normal cattle leads to greater potential of contamination of the surrounding environment and food supply [5]. Epidemiological studies of farms suggest that greater than 96% of O157 isolates originate from the 9% of animals that are super-shedders [6]. Each of the three principal components of the epidemiologic triad—the pathogen, the host, and the environment are likely to contribute to the SS phenotype, and several studies have investigated host and environmental factors associated with super-shedding [3, 6–9]. However, very little is known about the microbial factors that may contribute to super-shedding including the presence of virulence and adherence genes, the ability to form biofilm on biotic and abiotic surfaces, growth and survival rates, nutrient utilization, and other factors affecting overall fitness [4, 8, 9].

We have recently reported the sequence of the complete genomes of two SS strains and shown that SS strains display a distinctive adherence phenotype on rectal squamous epithelial cells recovered from the bovine RAJ [10, 11]. Our current comparative genomics investigations seek to define genomic features that may distinguish SS isolates from others in order to better understand the evolutionary history of SS stains and provide a framework for identifying molecular correlates associated with the SS phenotype in strains of O157.

Materials and methods

Bacterial strains

SS strains of O157 were isolated and characterized from approximately 3,500 cattle sampled during the summer months over a two-year period in U.S. Midwestern States, as described [12]. The isolates were characterized by phage typing and *Xba*I-based pulsed-field gel electrophoresis (PFGE) (Fig 1A, S1A Fig), and PCR was performed to confirm the presence of genes for O157 antigen, H7 flagella, γ-intimin, and at least one of the Shiga toxin genes as described [13]. Two representatives SS isolates, strains SS17 and SS52 were selected for the study [12]. Both SS17 and SS52 carried phage type 4 and were isolated from supershedding cattle at concentrations of 3.1×10^5 CFU/rectal swab and 6.8×10^5 CFU/rectal swab, respectively. The PFGE analysis of the SS52 matched the most common PFGE pattern of O157 isolates from human foodborne outbreaks [14]. *E. coli* O157 strains Sakai [NC_002695], EDL933 [NC_002655],
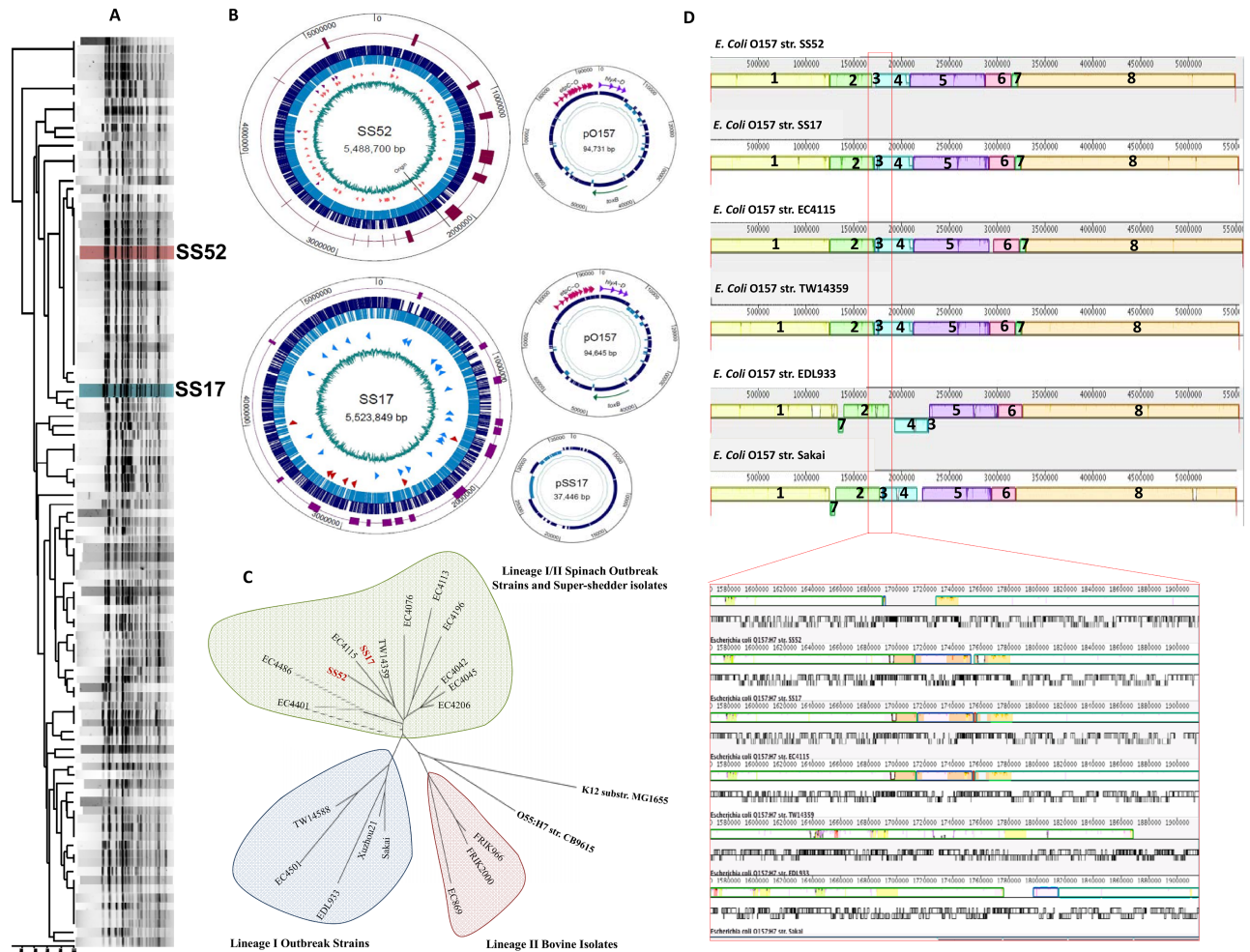
Fig 1. Comparative analysis of SS strains and reference O157 genomes. (A) Dendrogram representing cluster analysis of SS isolates based on PFGE patterns as previously described [12] highlighting the two representative isolates, SS17 and SS52, that have been completely sequenced; (B) Circular genome representations of SS17 and SS52. Larger circles depict chromosomal DNA with smaller circles representing plasmids. The blue circles represent the ORFs and the outer circle (purple) represents the phages in each genome; (C) Cladogram based on whole genome alignment reveals that SS17 and SS52 clusters closely with lineage I/II "spinach" outbreak isolates (EC4115 and TW14359) as compared with lineage I outbreak isolates (Sakai and EDL933) or the bovine lineage II isolates; (D) Whole genome alignments of SS strains with reference O157 strains using progressiveMauve depicting 8 homology blocks of similarity and patterns of divergence among the strains.

TW14359 [NC_013008], and EC4115 [NC_011353] were used as reference genomes for comparative analyses [15–18] (Table 1). All bacterial culturing were performed in a BSL-2 facility (IBC # 46638).

Genome sequencing and assembly

Genomic DNA was isolated using QIAGEN DNeasy Blood and Tissue Kit (cat. no. 69504) as previously described [10]. Genome sequencing and assembly procedures have previously been described for SS17 with minor modifications [11] [10]. Genomic DNA from SS52 was submitted to the Genomics Core Facility at The Pennsylvania State University for whole genome shotgun sequencing using the Ion Torrent PGM sequencer (Life Technologies, Grand Island, NY) [19], using a 318 sequencing chip and mate-pair sequencing. A total of 5.4M reads with

Table 1. Characteristics of *E. coli* O157:H7 isolates.

Isolate	Accession No.	Isolation Source	Outbreak Source	Reference
Sakai	NC_002695	Human	Radish Sprouts	Hayashi, *et al.* 2001
EDL933	NC_002655	Ground Beef	Hamburger	Perna, *et al.* 2001
TW14359	NC_013008	Human	Spinach	Kulasekara, *et al.* 2009
EC4115	NC_011353	Human	Spinach	Eppinger, *et al.* 2011
SS17	CP008805	Bovine Feces	SS Strain	Cote, *et al.* 2015
SS52	CP010304	Bovine Feces	SS Strain	This Study

https://doi.org/10.1371/journal.pone.0182940.t001

an average length of 169 bases were obtained with >168-fold coverage [11]. Whole genome restriction optical map was generated using *BamH*I digestion by OpGen, Inc. (Gaithersburg, MD) [20]. Both *de novo* and reference-guided assemblies were performed using DNASTAR SeqMan NGen 3.1 and Lasergene Suite V. 11.1 (Madison, WI) to obtain large contigs, and the genomes closed with a primer walking approach [11]. Circular representations of the genome were generated with GenVision 10 (SS17) and 11(SS52) (DNASTAR, Madison, WI) (Fig 1B).

Comparative genomics

Complete genomes of *E. coli* O157 strains Sakai [NC_002695], EDL933 [NC_002655], TW14359 [NC_013008], EC4115 [NC_011353], and the two SS strains, SS17 [AC_CP008805] and SS52 [AC-CP010304], were evaluated by comparative analyses [10, 11, 15–18]. Whole genome alignments were performed using Mauve and the progressiveMauve algorithm to identify single nucleotide polymorphisms (SNPs), genomic islands, and relative relationships between SS strains and reference strains [21, 22]. A lineage-specific polymorphism assay (LSPA) was performed using primers, described in Yang, *et. al* [23], and Sanger sequencing to confirm insertions and deletions. Circular representations of the genome were generated with GenVision 11 for SS52 (DNASTAR, Madison, WI). The whole genome phylogenetic tree was constructed using RealPhy [24] by inputting the Genbank accession IDs for the O157:H7 sequences. The tree builder used was PhyML and was run with 1016 bootstrap iterations (-b 1016), and using the output tree as the Maximum Likelihood tree. The tree was drawn with FigTree V1.4 (Institute of Evolutionary Biology, University of Edinburgh, http://tree.bio.ed.ac.uk/). The phage analysis was performed using PHAST software [25]. BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was performed for phage alignment including on a pairwise basis.

Results

Comparative genomic analysis of SS and reference isolates of O157

We have previously described the complete genome sequence of the prototypical SS strain, SS17 (accession no. CP008805 [10]), and recently completed the sequencing of a second O157 SS strain, SS52 (accession no. CP010304 [11], Fig 1B, S1B Fig). The availability of these high quality full genome sequences provided an opportunity to apply comparative genomic approaches to identify genetic features that may contribute to the SS phenotype.

The analyses showed that the genomes of SS17 and SS52 are similar to reference O157 strains [15–18] (Table 2). The whole genome phylogenetic and LSPA analyses reveal that both SS strains belong to O157 lineage I/II and cluster with isolates of spinach-associated outbreaks, including strains TW14359 and EC4115, and can be differentiated from the prototype lineage I O157 complete genomes of strains Sakai and EDL933, and bovine isolates lineage II (Fig 1C, S1C Fig) [12].

Table 2. Genome statistics of SS strains and reference O157 strains.

	SS17	SS52	EC4115	TW14359	Sakai	EDL933
Length of Sequence (Mb)	5.52	5.5	5.57	5.52	5.49	5.53
G+C ratio (%)	50.5	50.5	50.5	50.5	50.5	50.4
Coding DNA Sequences	5442	5655	5608	5555	5504	5587
Protein Coding Region (%)	88.5	88.8	86.8	86.9	86.9	86.5
Average ORF length (bp)	898	862	863	864	868	856
No. rRNA	22	22	22	22	22	22
No. tRNA	107	106	109	108	103	100
No. Plasmids	2	1	2	1	2	1

Alignments of the full genomes of SS17, SS52, and reference O157 strains using the progressiveMAUVE algorithm [21, 22] identified eight distinct blocks of homology, distinct prophage profiles, and two blocks with an inversion and dislocation (Fig 1D, S1D Fig).

Block 1 (Fig 1D, S1D Fig, highlighted in yellow) is a ~1.2Mb section present in all aligned genomes, except for the O157 strain EDL933, where it is 1.3Mb. The increase in the size is due to two insertions in EDL933 including a prophage that is inserted between the ATP dependent protease subunit gene, *clpA* and transposase genes and tRNA-SerW, and insertion of prophage CP-933M between *yccK*, a sulfur transfer protein and tRNA-SerT. The remaining block is well conserved, and begins with *thrA* gene, which is involved in the tryptophan biosynthetic process, and ends with *apg* gene, which plays a role in the late stage of autophagosome formation.

Block 2 (Fig 1D, S1D Fig, highlighted in light green) is a ~500kb region in strains SS17, SS52, EC4115, and TW14359. The area is about 20kb larger in EDL933, due to the insertion of mobile elements, and is rich in several prophages including CP-933N, CP-933O, and CP-933X. Also conserved in this area are genes including *cah*, a calcium-binding and heat-extractable autotransporter protein; *flgA* that is involved in the flagellar biosynthesis; and *fhuE*, an outer membrane receptor for ferric iron uptake. The block starts with *ycdG* gene (uracil permease) and ends with phage related proteins.

Block 3 (Fig 1D, S1D Fig, highlighted in dark blue) is a 2kb region in SS52, ~ 40kb in SS17, EC4115, and TW14359; and ~18 kb in both EDL933 and Sakai strains, and translocated and inverted EDL933 in comparison with the SS strains (SS17 and SS52) and the spinach outbreak strains (TW14359 and EC4115) strains. The primary difference in size results from a ~ 37.5kb insertion in SS17 that includes several transposases and integrases with sequence identity to phage CP-933O. This fragment in SS52 encodes for the putative Rem protein, putative cytoplasmic protein, and *rusA* gene, a Holliday Junction resolvase. The rest of the block is interrupted by phage and related proteins. The block in EC4115, and TW14359 contains genes including exonuclease VIII. In Sakai and EDL933, the area carries phage proteins, an integrase, *lomP*, an outer membrane protein gene and few other hypothetical proteins.

Block 4 (Fig 1D, S1D Fig, highlighted in light blue) is a ~350kb block in all genomes. Both blocks 3 and 4 have been interrupted by the insertion of phages CP-933O, and CP-933P, and the entire section consists of phage related proteins, and phage associated DNA methyl transferase. This block in SS52 contains several phage related proteins, intestinal colonization factors, phage regulators, phage antiterminator Q, non-LEE-encoded type III effector, and several transposases. This block is translocated and inverted in EDL933 genome. In Sakai, the translocation of areas 3 and 4, has led to a smaller section, which mostly contains phage related hypothetical proteins with the section including all phage- related proteins inserted in between both sections. EDL933 carries a similar section as Sakai with similar gene contents; however, area number 3 has translocated to block six.

Block 5 (Fig 1D, S1D Fig, highlighted in purple) is ~ 790kb in SS17, SS52, EC4115, and TW14359 strains, and 720kb in EDL933 and Sakai. This region is mostly conserved among SS17, SS52, EC4115, and TW14359. However, there are two (~5kb and ~2kb) insertions representing remnants of phages in both EDL933 and Sakai as compared to the SS isolates and other strains from lineage 1/II. The phage area contains phage CP-933R, other phages, phage—related proteins, and some transposase, which extend into the next homology block. The entire section is inverted in EDL933 as compared to the rest of the O157 genomes along with the translocation of the area three in other genomes.

Block 6 (Fig 1D, S1D Fig, highlighted in red) is ~ 272kb in SS17, SS52, EC4115, and TW14359 genomes, however, the area is smaller in EDL933 and Sakai, due to phage indels. Genome of EDL933 encodes for a ~ 25 kb CP-933V region that starts from this block with the *intV*, a integrase for prophage CP-933V after *yehV* (transcriptional regulator), and extended into block 5 ending with the gene *yehU* (component sensor protein). Shiga toxin gene stx1A is encoded within this region. This region encodes for many genes that are conserved among all O157 genomes. A few examples of genes that code for metabolism- related proteins in this region are *yoaE* (Putative transport), *edd* (Phosphogluconate dehydratase), *yeaG* (Protein kinase), or *topB* (Topoisomerase III). Gene *yeeJ*, encoding a hypothetical adhesin and an intimin/invasin homolog, is conserved among all genomes except for Sakai. This block also contains a gene encoding an ~4kb bacteriophage lambda -related protein and transposase in the genome of EC4115, not seen in the genomes of the other strains.

Block 7 (Fig 1D, S1D Fig, highlighted in dark green) is a 62kb section conserved among SS17, SS52, EC4115, and TW14359. The region is translocated to block 2 and inverted in both EDL933 and Sakai genomes. The region contains genes that encode for phages, Q933 antiterminator, and portal proteins of BP-933W. The region also contains genes encoding *stx2A2* and *stx2B* subunits in all genomes, and contains tRNA-Arg and prophage BP-933W-related proteins. In all genomes the block is flanked on one end with phage BP-933W integrase, encoding for the Shiga toxin II subunits A and B followed by protein Q protein of the bacteriophage BP-933W (*Q933*). Further analysis reveals that in both SS genomes, EC4115, and TW14359 genomes, the *stx2B* and *stx2A* subunits are followed by three tRNAs, tRNA-Arg, tRNA-Arg, and tRNA-Met, however, in Sakai and EDL933 genomes the three tRNAs are tRNA-Arg, tRNA-Arg, and tRNA-Ile.

Block 8 (Fig 1D, S1D Fig, highlighted in orange) is a 2.2Mb section that is well-conserved in SS17, SS52, EC4115, TW14359, EDL933 and Sakai, and due to the presence of phage and transposases, the block is ~ 35kb larger in Sakai.

Finally, the genomes of SS17, EC4115, and TW14359 encode for a small extra block not observed in SS52, EDL933, or Sakai (data not shown). The block is 821bp in SS17 located between block 5 and 6, and 1771bp located between blocks 3 and 4 in the EC4115 and TW14359 genomes, and encodes for hypothetical proteins.

Phages are the primary drivers of the recent evolutionary history of SS and other reference O157 isolates

The genomes of O157 (as well as, other *E. coli* strains) are naturally evolving by acquiring new genes, indels, rearrangements, point mutations, and phage-mediated lateral gene transfer. These acquisitions are recognized to be a major driver of the acquisition of new virulence or other traits that contribute to the fitness and adaptation of the pathogen in a particular environment [26–28]. Whole genome comparative analyses of phages in the SS and other reference O157 strains reveal between 17–19 phage regions in each of the genomes and provide strong evidence that phages are the primary drivers of genome scale variation in both SS and other
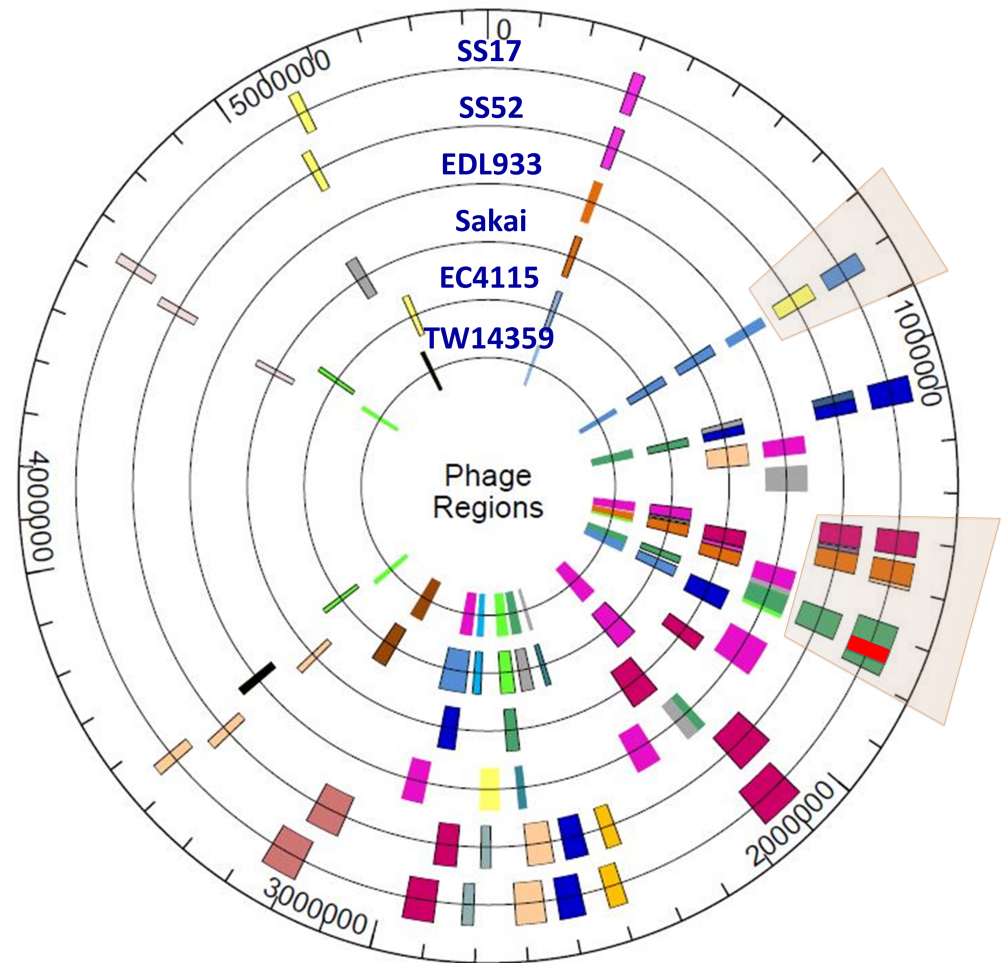
Fig 2. Phages in SS17, SS52, and reference O157 strains. The location and identities of phages were determined in each strain using the Phast server [25]. The color-coded blocks represent phage identity with similar colors showing the same phage on the circular genome. The highlighted area around 1.7MB depicts the insertion of a phage within another phage in SS17 but not SS52. From inner ring: TW14359, EC4115, Sakai, EDL933, SS52, and SS17.

https://doi.org/10.1371/journal.pone.0182940.g002

O157 strains (Fig 2; S1 and S2 Tables). SS17 and SS52 phage BLAST analysis indicates that SS17 has an insertion of ~ 37 Kb in the phage region starting from SS17-1718, position 1,726,958 to SS17-1774, position 1,765,158 (Fig 3). SS17, TW14359, and EC4115 share the same uninterrupted phage region, whereas SS52, EDL933, and Sakai have a different pattern. EDL933 and Sakai have a very similar pattern of phage insertion. The region starts with a transposase, and includes genes encoding for prophage CP-933O protein, integrase, and CP-933O encoded exodeoxyribonuclease VIII. It also encodes for antitermination Q and phage portal protein, and finally a terminase. In this region SS17 encodes for 140 CDS and SS52 70 CDS (S3–S5 Tables).

There are some similarities among the SS in this region as well. For instance, SS17, CD # 32 to 69 and SS52, CD # 14 to 51 carry the same pattern of phages, specifically in phage entero HK630_(NC-019723) (Fig 3; S5 Table).
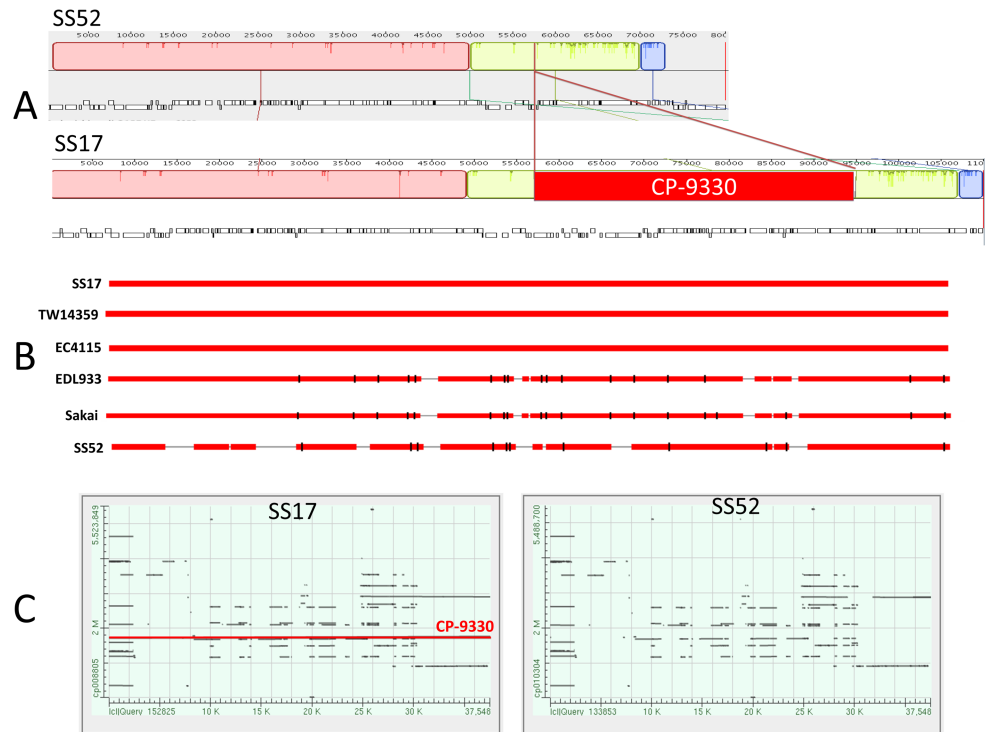
Fig 3. Phage diversity in SS and reference O157 strains. (A) MAUVE alignment of SS17 and SS52 showing the ~37 Kb insertion of CP-933O phage in the SS17 genome; (B) BLAST analysis of the region reveals that SS17, TW14359, and EC4115 all contain the same phage, whereas EDL933, Sakai, and SS52 do not; (C) Dot matrix representation of the results of megaBLAST analysis of the 37.5kb CP-933O region reveals its presence in SS17 (left panel) but not SS52 (right panel).

Analysis of SNPs in the genomes of SS and reference O157 strains

The analyses of SS52 revealed 801 SNPs in SS52 as compared with SS17, 384 SNPs as compared with TW14359, 604 as compared with EC4115, 2136 as compared with Sakai, and 3106 as compared with EDL933, not including ambiguous bases (Table 3). The analyses further identified a total of 167 non-synonymous SNPs (nsSNPs) between the core genomes of SS17 and SS52. Of interest are genes encoding for membrane proteins, such as *yjgN and yebS*; genes involved in the virulence and adhesion such as *eivA*, encoding a type III secretion protein, and a FimD-like (PapC) fimbriae anchoring protein, and *tccP*, the tir-cytoskeleton coupling protein; and interestingly genes involved in the iron acquisition and heme cycle, such as *chuA*, the

Table 3. Polymorphism analysis of SS52 compared with reference O157 genomes.

SNP category	SS17	EC4115	TW14359	Sakai	EDL933
nsSNP	167	152	114	679	866
sSNP	262	295	189	1009	1363
Nonsense	7	5	4	15	13
Homopolymer	139	26	7	57	113
Frameshift	129	18	9	60	143
Frameshift due to homopolymer	56	7	2	26	46
Total SNPs	801	604	384	2136	3106

outer membrane hemoglobin receptor, *chuS*, a heme transport protein, and *afuB*, a ferric iron ABC transporter (S6 Table).

Although phylogenetic analysis placed SS52 in the same lineage as EC4115 and TW14359, SS52 harbors several strain specific nsSNPs in genes as compared to TW14359 and EC4115 that are involved in the adherence, motility, and cell metabolism. For instance, when compared with TW14359, SS52 harbors nsSNPs in the genes encoding *fimB*, the type 1 fimbriae regulatory protein; *flgA*, encoding flagellar biosynthesis/ motility, and *fhuC*, a ferric hydroxamate transporter, involved in the transport of Fe3+ into O157 cells. As compared to EC4115, SS52 harbors nsSNPs in the genes encoding *flgA* (similar to TW14359), and *tar*, a chemotaxis protein involved in the flagellar regulon. Curiously, compared to EDL933 and SS17, SS52 has a nsSNP in the *tcc*P gene at two different sites. TccP is a tir-cytoskeleton coupling effector protein that plays a role in the production of the "attaching and effacing" (A/E) lesions in the host intestinal mucosa, the hallmark of the O157 pathogenesis and adherence in humans.

Discussion

Microbial factors leading to super-shedding in O157 are not fully understood. Our previous genomic and phenotypic comparisons of one SS isolate (SS17) showed that SS strains, including SS52, exhibit strong aggregative adherence patterns that are distinctive from the adherence pattern observed in reference O157 strain EDL933 [10]. The adherence characteristics of eight other isolates of SS, including SS52 also revealed the same strong aggregative adherence pattern on the RSE cells as well (S2 Fig) [10]. In order to identify whether there are shared genetic features that contribute to these observed distinctive adherence phenotype in SS strains, we have recently reported the sequence of a second SS isolate, SS52 [11], and have here performed a detailed comparative genomic analyses of SS17 and SS52 strains with other already characterized O157 genomes.

The results of our whole genome comparisons show that SS strains belong to lineage I/II and are clustered with O157 isolates associated with the "spinach" outbreaks of diseases first described in 2009 [17]. Further, our studies reveal that SS17 and SS52 are more closely related to EC4115 and TW14359 strains as compared with EDL933 and Sakai. Based on these analyses, it is tempting to speculate whether other lineage I/II isolates, including EC4115 and TW14359, may also have the potential to represent super-shedder isolates. However, given that the shedding status of the animal from which these two reference strains may have originated is unknown [15, 17], it will be interesting to test whether these two strains share the same adherence phenotype on bovine RAJ stratified squamous epithelial (RSE) that is characteristic of other SS strains [10].

The comparative genomic studies -show that even while the two supershedder isolates of SS17 and SS52 belong to the same genotypic class and display the same adherence phenotype, they are distinct at the genome level, with much of the genomic variation driven by phages and extrachromosomal elements. For instance, SS52 is missing the plasmid pSS17 that harbors a type IV secretion system and several hypothetical genes. This plasmid was likely either recently gained by SS17 or, conversely, lost by SS52 and thus may not have a direct role to play in maintenance of the SS phenotype. It will, however, be interesting in future studies to determine the presence of pSS17 amongst other SS isolates or cure SS17 of pSS17 to formally test its role in the hyper-adherent phenotype seen amongst SS strains.

The overall genomic comparison analyses of both SS genomes with reference genomes indicate that the length of sequences, G+C ratio and number of tRNAs in SS strains are, in general, similar to the reference strains. It is noteworthy that the average ORF length (at 898bp) is longer in the previously described SS17 genome than in reference strains [10]. In contrast, the

average ORF length of SS52 (862bp) is consistent with that of reference strains (average 863; range 856–868; Table 2), suggesting that the longer ORF length is particular to SS17 but perhaps not in other SS strains. Interestingly however, at ~ 88.7%, both SS17 and SS52 have a somewhat greater fraction of the genome that is protein coding as compared with the reference strains (average 86.7%; Table 2). While the biological significance of this apparently higher protein-coding region is unknown, we cannot rule out the possibility that this result is merely an artifact of the annotation process that may be resolved by updating the annotation of the reference genomes.

One of the more striking observations resulting from our investigation was the role that phages have played in diversifying the closely related SS isolates. Our studies provide strong evidence that the acquisition and loss of mobile genetic elements are likely the primary drivers of genomic change in SS (and other) O157 strains, and are responsible for the complex patterns of genome scale evolution in the recent evolutionary histories of these isolates. For instance, while SS17 and SS52 are clearly very closely related and share many genomic and phenotypic features including most of the phages, the presence of CP-9330 in SS17, TW14359 and EC4115 clearly differentiates this cluster from SS52 which lacks CP-9330 and is instead similar at this locus to the genomes of EDL933 and Sakai that are Lineage I strains and otherwise genetically and phenotypically distinct. While the role of CP-9330 in enhancing fitness or contributing to virulence of O157 is unknown, these observations suggest that it may not contribute to the SS phenotype, though this hypothesis needs to be formally tested. Together, the results suggest that the patterns of phage evolution drive much of the genome scale variation observed in SS isolates, and this is also evidenced by the distinctive PFGE patterns previously noted amongst SS isolates [12]. Given that phages appear to play such a prominent role in the diversification and evolution of O157 and other *E. coli*, it is tempting to speculate that they may play a role in development of the SS phenotype—perhaps through the transfer of a specific adaptive trait or regulatory mechanism that improves adherence or general fitness of the organism to replicate in the bovine intestinal tract. This hypothesis may be testable through additional comparative genomic characterization of phages amongst SS and other O157 isolates.

Finally, the results of our studies reveal several SNPs that differentiate SS strains from each other and from reference O157 isolates. Further, the patterns of distribution of these SNPs provide evidence for positive selection/adaptation of genes/traits involved in the virulence and adaptation of O157 to a life associated with a host—for instance, through shared non-synonymous substitutions in genes relating to adherence, iron acquisition, and motility. Given the distinctive hyper-adherent phenotype for RSE cells that is shared by SS strains, studies are needed to investigate the potential role of key virulence related loci identified to have variation through our current comparative genomics studies and how this may contribute to the SS phenotype and increased fitness of the organism.

In conclusion, our current comparative genomics studies of SS isolates have provided key insights into the overall genomic structure and patterns of evolution of SS strains of O157, as well as lay a strong foundation for future investigations on the microbial mechanisms that may drive super-shedding of this important food-borne pathogen.

Supporting information

S1 Table. PHAST analysis of the phage region in SS and reference strains.
(XLSX)

S2 Table. Summary of phage region covering areas of 800kb to 1.8Mb region.
(XLSX)

S3 Table. Phage encoded region 8 in SS17.
(XLSX)

S4 Table. Phage encoded CDS in SS52.
(XLSX)

S5 Table. Comparison of phage region 7 in SS17 with phage region 8 in SS52.
(XLSX)

S6 Table. nsSNP between SS52 and reference strains of O157 in genes related to virulence, adherence, and metabolism.
(XLSX)

S1 Fig. Comparative analysis of SS strains and reference O157 genomes. (A) Dendrogram representing cluster analysis of SS isolates based on PFGE patterns as previously described [12] highlighting the two representative isolates, SS17 and SS52, that have been completely sequenced; (B) Circular genome representations of SS17 and SS52. Larger circles depict chromosomal DNA with smaller circles representing plasmids. The blue circles represent the ORFs and the outer circle (purple) represents the phages in each genome; (C) Cladogram based on whole genome alignment reveals that SS17 and SS52 clusters closely with lineage I/II "spinach" outbreak isolates (EC4115 and TW14359) as compared with lineage I outbreak isolates (Sakai and EDL933) or the bovine lineage II isolates; (D) Whole genome alignments of SS strains with reference O157 strains using progressiveMauve depicting 8 homology blocks of similarity and patterns of divergence among the strains.
(TIF)

S2 Fig. Adherence patterns of eight additional SS strains of O157 (SS7, SS12, SS27, SS42, SS52, SS67, SS77, SS131) on RSE cells. The strong aggregative adherence is common to all tested supershedder strains. All slides are at 40x magnification, bacteria (green), with RSE cytokeratins (red), and nuclei of cells (blue).
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: Robab Katani, Rebecca Cote, Chitrita DebRoy, Terrance M. Arthur, Vivek Kapur.

Data curation: Robab Katani, Rebecca Cote, Indira T. Kudva, Terrance M. Arthur, Vivek Kapur.

Formal analysis: Robab Katani, Rebecca Cote, Vivek Kapur.

Investigation: Robab Katani, Vivek Kapur.

Methodology: Robab Katani, Indira T. Kudva, Chitrita DebRoy, Terrance M. Arthur.

Project administration: Vivek Kapur.

Resources: Robab Katani, Indira T. Kudva, Chitrita DebRoy, Terrance M. Arthur, Vivek Kapur.

Software: Robab Katani, Vivek Kapur.

Supervision: Vivek Kapur.

Visualization: Rebecca Cote, Vivek Kapur.

Writing – original draft: Robab Katani, Rebecca Cote, Indira T. Kudva, Vivek Kapur.

Writing – review & editing: Robab Katani, Rebecca Cote, Indira T. Kudva, Chitrita DebRoy, Terrance M. Arthur, Vivek Kapur.

References

1. Gyles CL. Shiga Toxin-Producing *Escherichia coli*: An Overview. J Anim Sci. 2007; 85(13):E45–62. https://doi.org/10.2527/jas.2006-508 PMID: 17085726

2. Mead PS, Slutsker L, Dietz V, McCaig LF, Bresee JS, Shapiro C, et al. Food-related illness and death in the United States. Emerg Infect Dis. 1999; 5(5):607–25. PMID: 10511517;

3. Arthur TM, Brichta-Harhay DM, Bosilevac JM, Kalchayanand N, Shackelford SD, Wheeler TL, et al. Super-shedding of *Escherichia coli* O157:H7 by Cattle and the Impact on Beef Carcass Contamination. Meat Science. 2010; 86(1):32–7. https://doi.org/10.1016/j.meatsci.2010.04.019 PMID: 20627603

4. Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ, Pearce MC, et al. Lymphoid Follicle-Dense Mucosa at the Terminal Rectum is the Principal Site of Colonization of Enterohemorrhagic *Escherichia coli* O157:H7 in the Bovine Host. Infect Immun. 2003; 71(3):1505–12. https://doi.org/10.1128/IAI.71.3.1505-1512.2003 PMID: 12595469

5. Matthews L, Low JC, Gally DL, Pearce MC, Mellor DJ, Heesterbeek JAP, et al. Heterogeneous Shedding of *Escherichia coli* O157 in Cattle and its Implications for Control. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(3):547–52. https://doi.org/10.1073/pnas.0503776103 PMID: 16407143

6. Omisakin F, MacRae M, Ogden ID, Strachan NJC. Concentration and Prevalence of *Escherichia coli* O157 in Cattle Feces at Slaughter. Appl Environ Microbiol. 2003; 69(5):2444–7. https://doi.org/10.1128/AEM.69.5.2444-2447.2003 PMID: 12732509

7. Carlson BA, Nightingale KK, Mason GL, Ruby JR, Choat WT, Loneragan GH, et al. Escherichia coli O157:H7 strains that persist in feedlot cattle are genetically related and demonstrate an enhanced ability to adhere to intestinal epithelial cells. Appl Environ Microbiol. 2009; 75(18):5927–37. Epub 2009/07/21. https://doi.org/10.1128/AEM.00972-09 PMID: 19617387;

8. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. Super-shedding and the Link Between Human Infection and Livestock Carriage of *Escherichia coli* O157. Nat Rev Micro. 2008; 6(12):904–12.

9. Chase-Topping ME, McKendrick IJ, Pearce MC, MacDonald P, Matthews L, Halliday J, et al. Risk Factors for the Presence of High-Level Shedders of *Escherichia coli* O157 on Scottish Farms. J Clin Microbiol. 2007; 45(5):1594–603. https://doi.org/10.1128/JCM.01690-06 PMID: 17360845

10. Cote R, Katani R, Moreau MR, Kudva IT, Arthur TM, DebRoy C, et al. Comparative Analysis of Super-Shedder Strains of *Escherichia coli* O157:H7 Reveals Distinctive Genomic Features and a Strongly Aggregative Adherent Phenotype on Bovine Rectoanal Junction Squamous Epithelial Cells. PLoS One. 2015; 10(2):e0116743. https://doi.org/10.1371/journal.pone.0116743 PMID: 25664460.

11. Katani R, Cote R, Raygoza Garay JA, Li L, Arthur TM, DebRoy C, et al. Complete Genome Sequence of SS52, a Strain of *Escherichia coli* O157:H7 Recovered from Supershedder Cattle. Genome Announcements. 2015; 3(2). https://doi.org/10.1128/genomeA.01569-14 PMID: 25792068

12. Arthur TM, Ahmed R, Chase-Topping M, Kalchayanand N, Schmidt JW, Bono JL. Characterization of *Escherichia coli* O157:H7 strains isolated from supershedding cattle. Appl Environ Microbiol. 2013; 79 (14):4294–303. Epub 2013/05/07. https://doi.org/10.1128/AEM.00846-13 PMID: 23645203;

13. Arthur TM, Nou X, Kalchayanand N, Bosilevac JM, Wheeler T, Koohmaraie M. Survival of *Escherichia coli* O157:H7 on Cattle Hides. Applied and Environmental Microbiology. 2011; 77(9):3002–8. https://doi.org/10.1128/AEM.02238-10 PMID: 21398483

14. PulseNet. Centers for Disease Control and Prevention-PulseNet Centers for Disease Control and Prevention2013. http://www.cdc.gov/pulsenet/.

15. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic Anatomy of *Escherichia coli* O157:H7 Outbreaks. Proceedings of the National Academy of Sciences. 2011; 108(50):20142–7. https://doi.org/10.1073/pnas.1107176108 PMID: 22135463

16. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, et al. Complete Genome Sequence of Enterohemorrhagic *Eschelichia coli* O157:H7 and Genomic Comparison with a Laboratory Strain K-12. DNA Research. 2001; 8(1):11–22. https://doi.org/10.1093/dnares/8.1.11 PMID: 11258796

17. Kulasekara BR, Jacobs M, Zhou Y, Wu Z, Sims E, Saenphimmachak C, et al. Analysis of the Genome of the *Escherichia coli* O157:H7 2006 Spinach-Associated Outbreak Isolate Indicates Candidate Genes That May Enhance Virulence. Infect Immun. 2009; 77(9):3713–21. https://doi.org/10.1128/IAI.00198-09 PMID: 19564389

18. Perna NT, Plunkett G, Burland V, Mau B, Glasner JD, Rose DJ, et al. Genome Sequence of Enterohaemorrhagic *Escherichia coli* O157:H7. Nature. 2001; 409(6819):529–33. https://doi.org/10.1038/35054089 PMID: 11206551

19. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An Integrated Semiconductor Device Enabling Non-Optical Genome Sequencing. Nature. 2011; 475(7356):348–52. https://doi.org/10.1038/nature10242 PMID: 21776081

20. Latreille P, Norton S, Goldman B, Henkhaus J, Miller N, Barbazuk B, et al. Optical Mapping as a Routine Tool for Bacterial Genome Sequence Finishing. BMC Genomics. 2007; 8(1):321. https://doi.org/10.1186/1471-2164-8-321 PMID: 17868451

21. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements. Genome Res. 2004; 14(7):1394–403. Epub 2004/07/03. https://doi.org/10.1101/gr.2289704 PMID: 15231754;

22. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. PLoS ONE. 2010; 5(6):e11147. https://doi.org/10.1371/journal.pone.0011147 PMID: 20593022

23. Yang Z, Kovar J, Kim J, Nietfeldt J, Smith DR, Moxley RA, et al. Identification of Common Subpopulations of Non-Sorbitol-Fermenting, b-Glucuronidase-Negative *Escherichia coli* O157:H7 from Bovine Production Environments and Human Clinical Samples Applied and Environmental Microbiology. 2004; 70(11):6846–54. https://doi.org/10.1128/AEM.70.11.6846-6854.2004 PMID: 15528552

24. Bertels F, Silander OK, Pachkov M, Rainey PB, van Nimwegen E. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Molecular biology and evolution. 2014; 31(5):1077–88. https://doi.org/10.1093/molbev/msu088 PMID: 24600054;

25. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic acids research. 2011; 39(Web Server issue):W347–52. https://doi.org/10.1093/nar/gkr485 PMID: 21672955;

26. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of E. coli commensal and pathogenic isolates. J Bacteriol. 2008; 190(20):6881–93. https://doi.org/10.1128/JB.00619-08 PMID: 18676672;

27. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS genetics. 2009; 5(1):e1000344. https://doi.org/10.1371/journal.pgen.1000344 PMID: 19165319;

28. Zhang Y, Laing C, Steele M, Ziebell K, Johnson R, Benson AK, et al. Genome evolution in major *Escherichia coli* O157:H7 lineages. BMC Genomics. 2007; 8:121. https://doi.org/10.1186/1471-2164-8-121 PMID: 17506902;