

RESEARCH ARTICLE

Whole genome shotgun sequencing revealed highly polymorphic genome regions and genes in *Escherichia coli* O157:H7 isolates collected from a single feedlot

Xin Gao¹✉, Xun Yang¹✉, Lance Noll², Xiaorong Shi², Jay Worley¹, Marc Allard³, Eric Brown³, T. G. Nagaraja², Jianghong Meng¹*

1 Joint Institute for Food Safety and Applied Nutrition, Center for Food Safety and Security Systems, University of Maryland, College Park, MD, United States of America, **2** Department of Diagnostic Medicine/Pathobiology, College of Veterinary Medicine, Kansas State University, Manhattan, KS, United States of America, **3** Center for Food Safety and Applied Nutrition, US Food and Drug Administration, College Park, MD, United States of America

✉ These authors contributed equally to this work.

* jmeng@umd.edu



OPEN ACCESS

Citation: Gao X, Yang X, Noll L, Shi X, Worley J, Allard M, et al. (2018) Whole genome shotgun sequencing revealed highly polymorphic genome regions and genes in *Escherichia coli* O157:H7 isolates collected from a single feedlot. PLoS ONE 13(8): e0202775. <https://doi.org/10.1371/journal.pone.0202775>

Editor: Chitrira DebRoy, The Pennsylvania State University, UNITED STATES

Received: May 1, 2018

Accepted: August 8, 2018

Published: August 28, 2018

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: Individual sequence data are available from the NCBI Sequence Read Archive (SRA) under project number PRJNA437074.

Funding: This material is based upon work that is supported by the National Institute of Food and Agriculture (<https://nifa.usda.gov/>), U. S. Department of Agriculture, under award number 2012-68003-30155 and in part by the Joint Institute for Food Safety and Applied Nutrition

Abstract

Escherichia coli serotype O157:H7 continues to pose a serious health threat to human beings. Cattle, a major reservoir of the pathogen, harbor *E. coli* O157:H7 in their gastrointestinal tract and shed variable concentrations of *E. coli* O157:H7 into the environment. Genetic characterization of cattle-shed *E. coli* O157 strains is of interest to the livestock industry, food business, and public health community. The present study applied whole genome shotgun sequencing (WGS) and single nucleotide variant (SNV) calling to characterize 279 cattle-shed *E. coli* O157:H7 strains isolated from a single feedlot located in southwestern region of the US. More than 4,000 SNVs were identified among the strains and the resultant phylogenomic tree revealed three major groups. Using the Sakai strain genome as reference, more than 2,000 SNVs were annotated and a detailed SNV map generated. Results clearly revealed highly polymorphic loci along the *E. coli* O157:H7 genome that aligned with the prophage regions and highly variant genes involved in processing bacterial genetic information. The WGS data were further profiled against a comprehensive virulence factor database (VFDB) for virulence gene identification. Among the total 285 virulence genes identified, only 132 were present in all the strains. There were six virulence genes unique to single isolates. Our findings suggested that the genome variations of the *E. coli* O157:H7 were mainly attributable to dynamics of certain phages, and the bacterial strains have variable virulence gene profiles, even though they came from a single cattle population, which may explain the differences in pathogenicity, host prevalence, and transmissibility by *E. coli* O157:H7.

(JIFSAN), University of Maryland (<http://jifsan.umd.edu/>).

Competing interests: The authors have declared that no competing interests exist.

Introduction

Escherichia coli serotype O157:H7 is one of the most important foodborne pathogens and poses serious threat to human health, causing an estimated 73,480 illnesses, 2,168 hospitalizations and 61 deaths within the United States alone each year [1]. USDA/ERS estimated the annual cost of the *E. coli* O157:H7 cases at \$271.4 million, an enormous disease burden that cannot be ignored (<https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses/>).

E. coli O157:H7 produces Shiga toxins and expresses type III secretion systems, which lead to a wide spectrum of diseases including the severe ones such as bloody diarrhea, hemolytic uremic syndrome (HUS), and even death [2, 3]. Infectious dose of the *E. coli* O157:H7 is remarkably low—as few as 10–100 colony-forming units (CFUs) may induce symptoms, and as a result, a zero tolerance policy against *E. coli* O157:H7 is enforced in the processed ground beef [4, 5]. Various transmission routes of *E. coli* O157:H7 have been proposed, and the primary one is presumably via consumption of contaminated food or water. However, in many outbreaks, the transmission details simply remained unclear [1].

Since its first report in 1982 during an investigation of hemorrhagic colitis [6, 7], *E. coli* O157:H7 outbreaks in the US have been repeatedly linked to raw or undercooked food of bovine origin [8–11]. Food industry has carried out extensive work to track *E. coli* O157:H7 along the whole beef production chain, and observed that cattle actually may harbor *E. coli* O157:H7 in their gastrointestinal tract but without presenting any clinical manifestations [12, 13] and the cattle fecal shedding contains viable, infectious *E. coli* O157:H7 [14, 15]. Further epidemiology studies showed that up to 30% of the cattle would be such asymptomatic *E. coli* O157:H7 carriers [16, 17], which makes cattle the most important *E. coli* O157:H7 reservoir for the food and water contaminations. Unfortunately, few intervention measures appear effective in reduction of *E. coli* O157:H7 in cattle population and their immediate environments [18].

Characterization of the cattle-shed *E. coli* O157:H7 isolates, particularly their genome variability, transmission modes, and contamination mechanisms, is always necessitated for the interests of food safety-related trace-back investigations, risk assessment practice, and regulatory framework decision. Traditionally, *E. coli* O157:H7 is characterized utilizing a range of phenotypic- and/or genotypic-based typing techniques [19–23], which have significantly facilitated *E. coli* O157:H7 detection, clonal characterization, serotype determination, lineages differentiation, and virulence assessment [24–27]; however, the processes are generally time-consuming, expensive, and less sensitive. The advances in next-generation sequencing in the past decade have made it possible to perform whole genome sequencing (WGS) of any organism including *E. coli* O157:H7 rapidly and at affordable costs [28, 29]. More importantly, mining the WGS data allows simultaneous examination of multiple groups of featured gene markers, and consequently has a discriminatory power superior to any of the existing traditional methods [30, 31]. There is an ever-increasing body of evidence demonstrating that WGS has become a dominant tool in foodborne pathogen and infection investigations [32].

WGS studies on the *E. coli* O157:H7 have been carried out, but mostly for prevalence investigations and usually small sample collections. In the present study, we applied WGS and single nucleotide variant (SNV) analysis to 279 cattle-shed *E. coli* O157:H7 strains and focused on their genetic diversity and the subsequent effects. The strains were isolated from cattle feces collected from a single feedlot in southwestern region of the US. This was a large single-location collection of *E. coli* O157:H7. We analyzed those strains together with the reference *E. coli* O157:H7 Sakai genome, and other *E. coli* O157:H7 assemblies available from the GenBank to explore a deep phylogenetic relationship among those strains. The identified SNVs were

annotated to the reference *E. coli* O157:H7 Sakai genome and the highly variable genomic regions and gene-set were revealed. The WGS data were further profiled against a comprehensive virulence factor database which contained more than 26,000 records, for virulence gene identification. Our results presented valuable information and novel insights for understanding the *E. coli* O157:H7 genome and virulence genotype evolution.

Methods

Bacterial strains

Escherichia coli strains included in the study were isolated from feces collected from 576 cattle housed in 48 soil-surfaced pens (40 ft. × 115 ft. with 36 ft. of bunk space) in a feedlot of South Western US between July and October of 2015. The Institutional Animal Care and Use Committee at Kansas State University approved the research for sample collection and detection and isolation of strains that were used in the study (IACUC # 3674). All the fecal samples were collected with a plastic spoon from freshly-defecated pen-floor fecal pats with care taken to avoid ground contamination. The spoon with feces was placed into a Whirl-pak bag (Nasco, Ft. Atkinson, WI), then transferred in cold storage to the Kansas State University Pre-harvest Food Safety Laboratory.

Fecal samples were individually processed in the lab, with 2 g of the fecal sample mixed well with the 18 ml *E. coli* broth (EC; Difco, ThermoFisher, Waltham, MA) and then incubated at 40°C for 6 hrs. An aliquot of the enriched fecal suspensions was subjected to immunomagnetic separation (IMS) procedure with O157 IMS beads (Abraxis®, Warminster, PA). Twenty-five microliters of the O157 bead suspensions were spread-plated onto sorbitol MacConkey with cefixime and tellurite (CT-SMAC) medium, and the plates were incubated at 37°C for 20–24 hours. Up to six sorbitol-negative colonies from CT-SMAC were randomly picked, inoculated individually onto blood agar plates (Remel), and then incubated at 37°C for 24 hours. Isolates obtained from the CT-SMAC plate were tested for agglutination for O157 antigen by latex agglutination, and if positive, tested for indole production. Isolates positive for agglutination and indole were tested by a six-plex PCR assay that targeted *rfbE*_{O157}, *fliC*_{H7}, *eae*, *stx1*, *stx2* and *ehxA* genes [33]. Confirmation of *E. coli* O157:H7 was based on a gene profile positive for *rfbE*_{O157}, *eae*, *fliC*_{H7} and at least one Shiga-toxin gene (either *stx1* or *stx2*). Confirmed isolates were stored in CryoCare™ beads (Key Scientific Products, Round Rock, TX).

Preparation of genomic DNA for high-throughput whole genome sequencing

E. coli O157:H7 strains were re-streaked onto blood agar and then subcultured in Tryptone Soy broth. The genomic DNA was extracted from the broth culture using DNeasy Blood and Tissue Kit with the QIAcube robotic workstation (Qiagen, Germantown, MD). The multiplexed genomic libraries were constructed using Nextera XT DNA Library Preparation Kit and MiSeq Reagent Kits v2 (500 Cycles, Illumina, Inc.) following the manufacturer's instruction. Whole genome shotgun sequencing was performed using an Illumina MiSeq platform (Illumina, San Diego, CA) to generate paired-end reads.

Sequence analysis

Raw sequencing data were subjected to *De novo* genome assembly using SPAdes 3.6.0 [34] with the default settings, and strains with spurious assembly size removed. The raw reads data from the remaining strains were first processed using the *in-house* script wrapper to remove

the reads of low quality (base score cutoff: 20), low complexity (dust score cutoff: 7) and short length (length cutoff: 150) [35, 36].

The raw-read sequencing files with a calculated sequencing depth greater than 20X from the processed data were retained for the SNV call using the kSNP3 algorithm [37] and phylogenetic tree construction using FastTree (included in the kSNP3 package). SNV annotation was performed using the SnpEff script [38]. The clade classification followed the Manning system [39] using the BLAT aligner with 100% identity [39, 40]. The virulence profiling was performed using the biobakery-ShortBRED package [41] against the downloaded Virulence Factors Database (VFDB) peptide sequences (www.mgc.ac.cn/Vfs) [42], which contained more than 26,000 records from various pathogen sources. All the statistical graphs were created in R.

Dataset accession

Individual sequence data are available at the SRA, which is hosted by the NCBI, under project number PRJNA437074. Review link to the metadata: ftp://ftp-trace.ncbi.nlm.nih.gov/sra/review/SRP134008_20180307_125217_37d5c0b6b354bc3c790d2696b42756c9. Individual sequence data are available at GenBank with accession number: SRP134008.

Results

SNV analysis and SNV-based Phylogenomic structure characterization

The 279 *E. coli* O157:H7 strains used in this study originated from the fecal samples of 525 crossbred cattle housed in a single feedlot. All the sequencing data passed our *in-house* quality criteria. To decipher the high-resolution phylogenomic structure of the strains, we applied the WGS, followed by k-mer based SNV analysis [37]. A fully closed genome of *E. coli* O157:H7 Sakai strain (GCA_000008865.1_ASM886v1) was downloaded from the GenBank and included in the study for comparison and SNV annotation. The analysis called a total of 4,047 SNVs (1,798 to 2,406 SNVs for individual strains, [S1 Table](#)). Interestingly, none of the SNVs were present in all the analyzed genomes. Absence of the common SNVs, across all of the analyzed strains (core SNVs) indicated that indel events might be important during the evolution of the *E. coli* O157:H7 genome.

With the Sakai strain genome included, we found that 2,274 SNVs (56.2%) were derived from the k-mers that were uniquely mapped to this reference, 126 SNVs (3.1%) from the k-mers mapped twice, and 1,648 SNVs (40.7%) from the k-mers not mapped at all. Such high percentage of the unmapped SNV-containing k-mers suggested that our *E. coli* O157:H7 collection was quite divergent from the reference Sakai strain. The identified SNVs were used for phylogenomic tree construction ([Fig 1](#), [S1 Text](#)). The resultant tree topology revealed 3 distinct clusters, which roughly correspond to groups 6, 7, and 8, defined by Manning Clade system ([S2 Table](#)) [39]. The clade 6 strains accounted for the largest number of the total collection ($n = 191$, 68.5%), followed by those belonging to clade 7 ($n = 50$, 17.9%) and clade 8 ($n = 35$, 12.5%), respectively. We also found that three strains (KSU151, KSU155, and KSU247) that could not be assigned to any clade due to the absence of the clade-specific SNVs, but clustered well with the remaining clade 6 strains. Two clade 6 strains KSU263/KSU264 appeared as outliers on the phylogeny tree, as evidenced by a separate cluster with a long branch in [Fig 1](#).

To put our isolate collection into more general genetic context, an additional 175 genomes related to the *E. coli* O157:H7, including those either contained O157 (*E. coli* O157:X) or H7 (*E. coli* X:H7) ([S3 Table](#)), as well as one non-pathogenic *E. coli* K12 (GCA_000981485.1_Ecoli-K12AG100), were downloaded from the GenBank and subjected to another round of SNV call and tree construction ([Fig 2](#), [S2 Text](#)). With the *E. coli* K12 genome used as an outgroup for

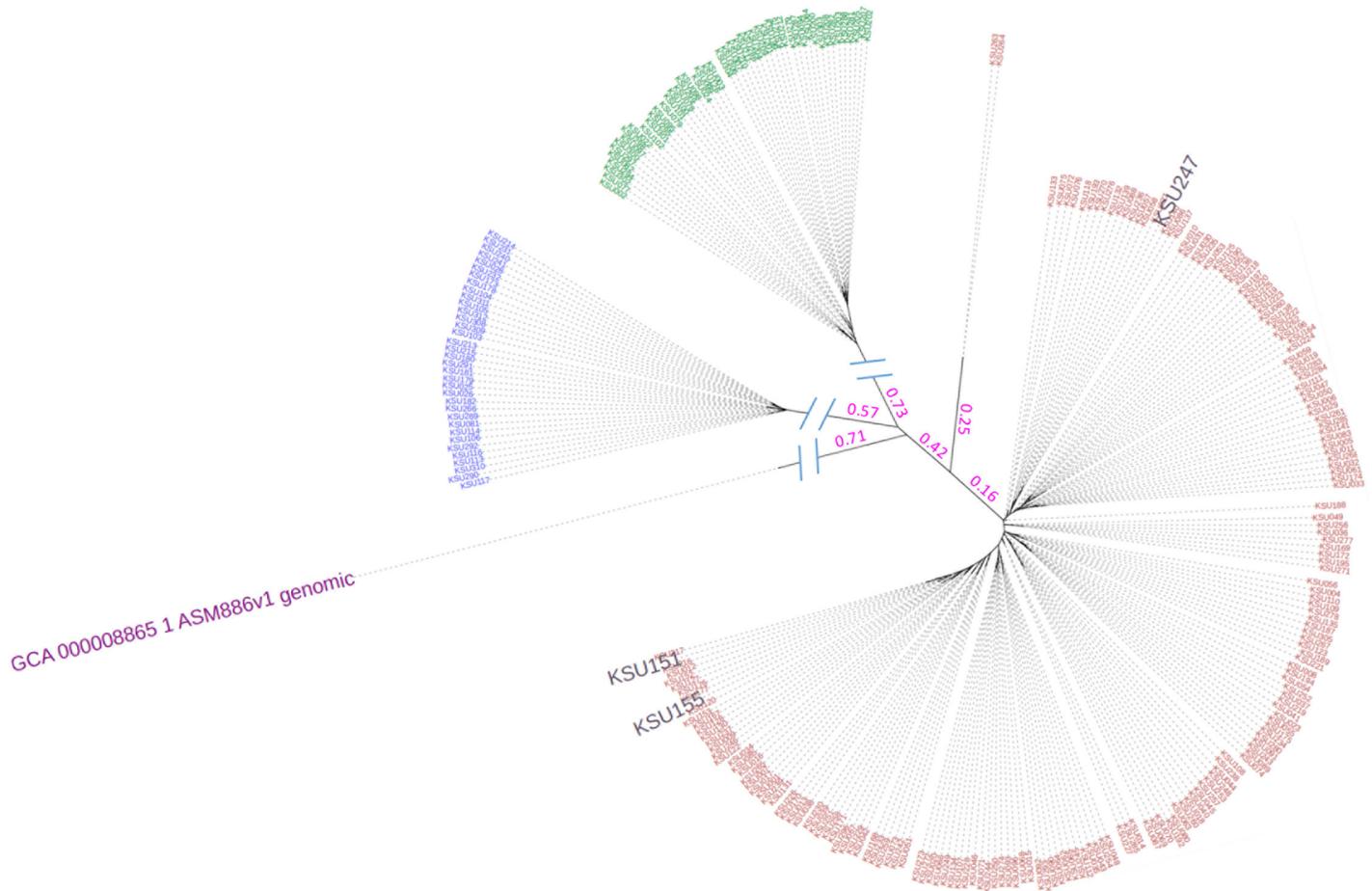


Fig 1. Phylogenomic relatedness of the *E. coli* O157 strains isolated from cattle feces from a single feedlot. ML phylogenetic tree; red, strains in clade 6; green, strains in clade 7; blue, strains in clade 8; rosy-brown, strains that could not be assigned to a clade; and purple, *E. coli* O157:H7 Sakai strain for reference.

<https://doi.org/10.1371/journal.pone.0202775.g001>

the phylogeny, data demonstrated that all the *E. coli* O157:H7 strains tended to cluster together, consistent with the commonly accepted clonal theory of the *E. coli* O157:H7 population. Our outlier strains, KSU263/KSU264, stayed within the clonal *E. coli* O157:H7 strain cluster, but more closely related to the 11 *E. coli* O157:H7 strain 2011EL, all of which were clinical isolates (<https://www.ncbi.nlm.nih.gov/pathogens>) [43]. The data, also unsurprisingly, showed that *E. coli* O55:H7 strains were closely related to the *E. coli* O157:H7.

SNV annotations

Since about 60% (56.2% + 3.1%) of the identified SNVs were mapped to the Sakai genome, we further performed SNV annotation to better understand how the SNVs would affect the biology of O157:H7 *E. coli* strains. We only focused on the SNVs that were derived from unique k-mers ($n = 2,274$) to avoid the interference of possible duplicated/repetitive genome regions. The SNVs occurred in both bacterial chromosome ($n = 2,226$) and bacterial plasmids ($n = 48$) (S4 Table). Among the 2,274 SNVs, 1,915 were located within the coding region including 13 that had impact on two genes due to the gene overlapping on the Sakai genome, and 359 within the intergenic region. The SNVs that were annotated into the coding regions comprised

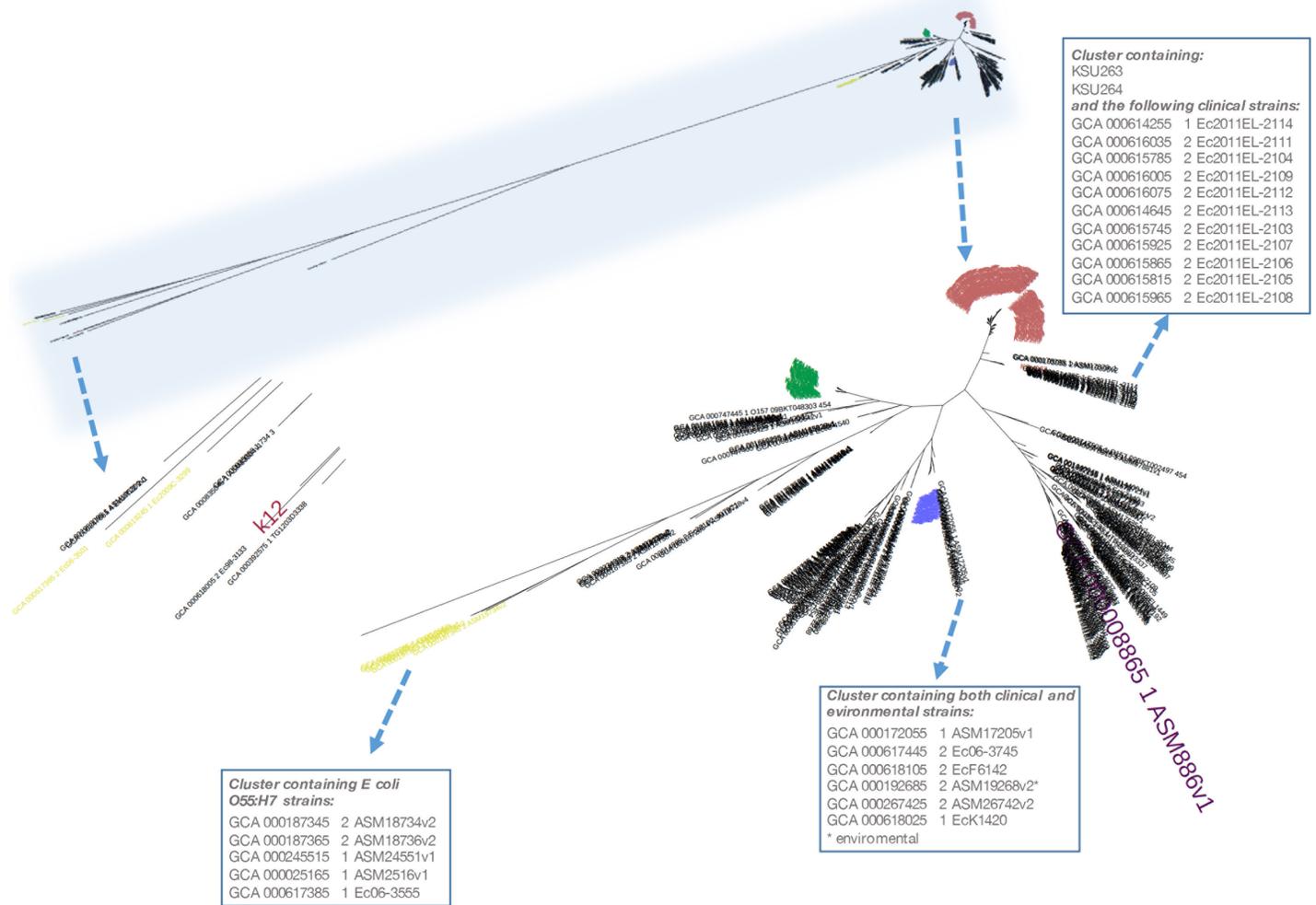


Fig 2. Phylogenomic relatedness among the *E. coli* O157 strains isolated from cattle feces from a single feedlot and the relevant genomes downloaded from the GenBank. Black, the downloaded genomes; red, strains in clade 6; green, strains in clade 7; blue, strains in clade 8; yellow, *E. coli* -H7, including *E. coli* O55:H7; purple, *E. coli* O157:H7 Sakai strain; and dark red, *E. coli* K12. Boxes contains the genome accession numbers corresponding to a given cluster.

<https://doi.org/10.1371/journal.pone.0202775.g002>

of 937 synonymous changes, 973 nonsynonymous changes, and 10 synonymous and nonsynonymous changes (dual SNVs).

The large strain collection of the present study allowed us to identify and annotate in detail a considerably higher number of SNVs compared with what has been reported before [44–47]. Using the Sakai genome, we were able to estimate that about 79% (1,796/2,274) of the SNVs had allele frequencies (AF) greater than 1%, suggesting that most of the SNVs were evolutionarily stable. SNVs were further grouped according to their AFs, and the results (S1 Fig) showed that the majority of the SNVs had AF less than 20% (n = 1,547, 68%), followed by those with AF between 60–70% (n = 243, 11%).

Fig 3 showed all the SNVs with AF greater than 1% along the Sakai genome. The track of the coding-region SNVs distinctively revealed several variant hot-spots (more than 15 SNVs within a genomic window of 2,000 bases versus an average of less than 1 SNV per 2,000-base in the Sakai genome window), as shown in Fig 4. All of the hot-spots were perfectly aligned in the prophage regions. For instance, genome regions around 1.25 M, 1.6M and 1.9 M

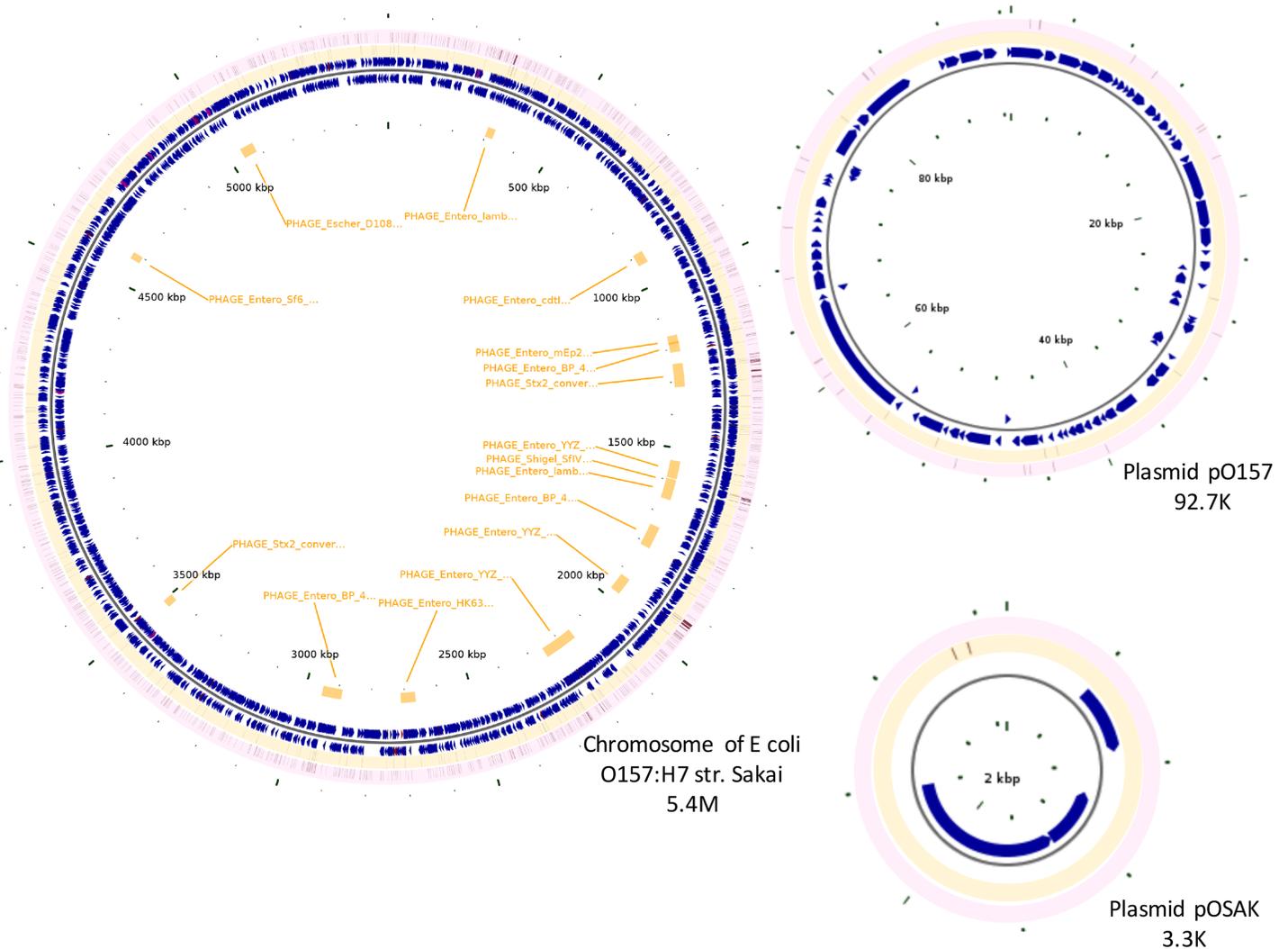


Fig 3. Single nucleotide variants on *E. coli* O157:H7 genome. Pink circular track, SNVs in genes; orange circular track, SNVs in intergenic regions; orange bars, phage regions; and blue arrows, genes.

<https://doi.org/10.1371/journal.pone.0202775.g003>

corresponded to the prophage *stx_2*, prophage *Shigel_SfIV*, and prophage *entero_YYZ* loci, respectively. Instead, the distribution of the intergenic SNVs tended to be more stochastic, possibly because of overall low count ($n = 359$, as indicated above) of the SNVs annotated to such regions. Several intergenic regions enriched with the SNVs appeared to align within hot-spots in the coding regions.

SNVs were also annotated to the genome of non-pathogenic *E. coli* K12 strain AG100. When K12 was used as the genomic reference, a total of 143,463 SNVs were identified, and 92,221 of them uniquely annotated to the K12 genome. Such high number of total SNVs, when compared with that in the background of Sakai, is obviously consistent with the prominent genetic divergence between the pathogenic *E. coli* O157:H7 and non-pathogenic *E. coli* K12. It is worth noting that more than 28,000 SNVs (>35% of the annotated SNVs) presented allele frequencies greater than 0.7 as shown in (S2A Fig), strongly supporting that those genomic variants are fixed or nearly fixed within the *E. coli* O157:H7 genome. Comparison to the *E. coli* K12 strain did not reveal SNV hot-spots, indicating that divergence between the *E. coli* O157:

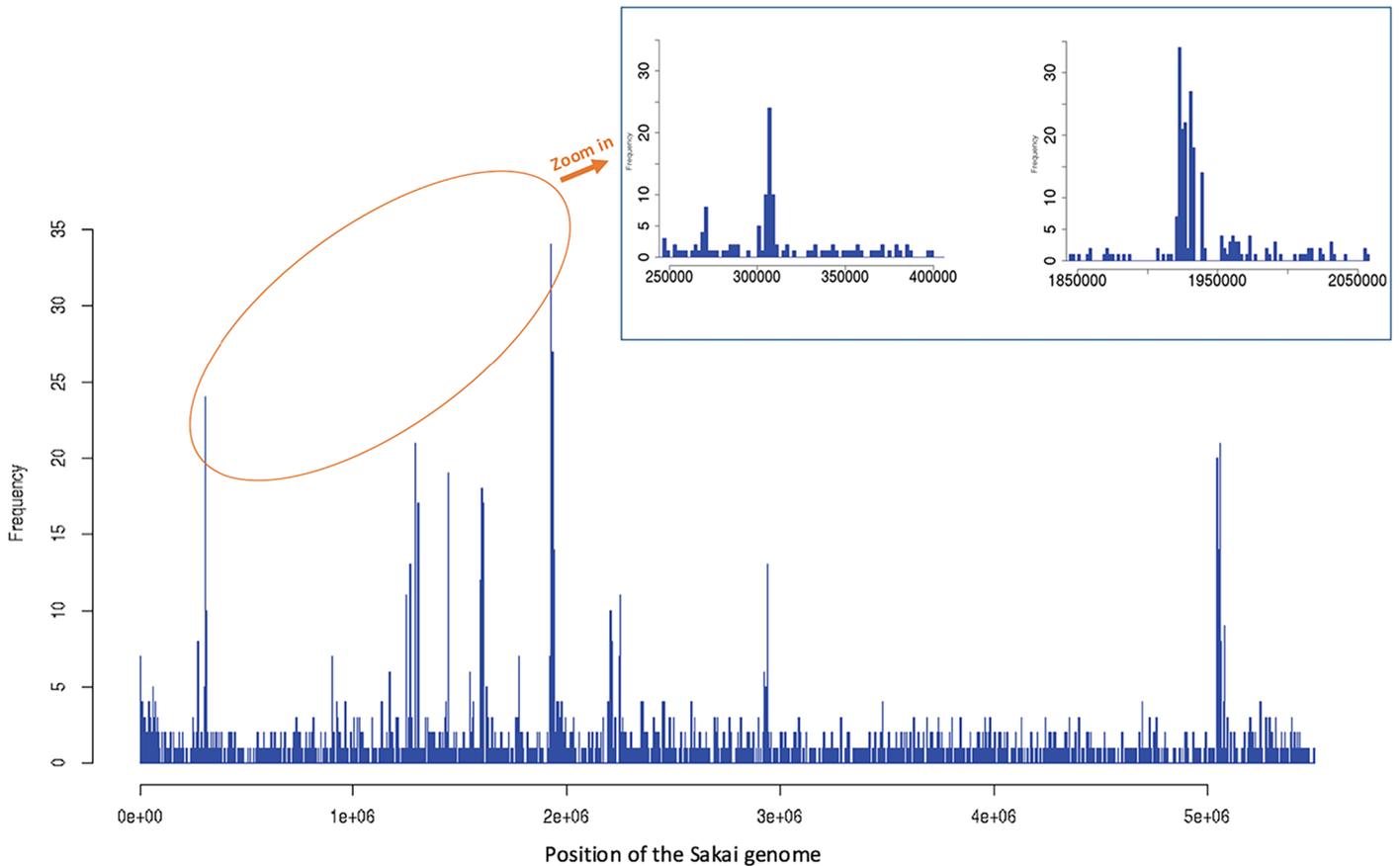


Fig 4. Number of single nucleotide variants (SNV) per 2,000 bases along the *E. coli* Sakai genome. SNV-rich regions were zoomed in.

<https://doi.org/10.1371/journal.pone.0202775.g004>

H7 and *E. coli* K12 has no preference on specific genetic background. Even though high numbers of SNVs were found when compared with the K12 genome, short genomic stretches (genome positions: 0.3 M and 2.8 M) were SNV free and the corresponding genes may possess important house-keeping functions for bacterial survival (S2B Fig).

Functional annotation for the genes associated with the SNVs

High-quality, deep genome sequencing in this study was key for positioning SNVs into individual *E. coli* O157:H7 Sakai strain genes. Out of 5,584 Sakai genes, 1,201 contained SNVs with the numbers ranging from 1 to 26, including 1,193 chromosomal and 8 plasmid genes (Fig 3, S4 Table). The SNV number for a given gene was then normalized against its length to obtain the corresponding SNV density (S5 Table). We found that 15 coding sequences possessed more than 15 SNVs per 1,000 bases (SPK, number of SNVs per 1K genome) (Table 1). Those genes were considered to be enriched for SNVs given that the genome-wise SPK was less than 1. Further analysis showed that the SNV-enriched genes were generally shorter with a mean length of only 435 bp (compared with the mean length of about 1 kb of the overall Sakai genes), and all were located on chromosomal DNA. A majority (9 out of 15, 60%) of the SNV-enriched genes coded for hypothetical proteins without specific functions assigned, and the ones that did have functional annotations (n = 6) were all involved in genetic information

Table 1. Genes enriched for SNVs from *E. coli* O157:H7.

Gene ID	Description	Length	SNV density (Counts/1000 bases)
ECs0281	hypothetical protein	593	16.86
ECs0284	DNA-invertase	554	18.05
ECs1395	hypothetical protein	872	17.20
ECs1577	Icd-like protein	197	20.30
ECs1581	hypothetical protein	299	20.07
ECs1587	putative single stranded DNA-binding protein	410	21.95
ECs1588	putative transcriptional activator	272	18.38
ECs1599	hypothetical protein	281	35.59
ECs1931	hypothetical protein	188	31.91
ECs1932	restriction alleviation and modification enhancement protein	149	20.13
ECs1933	recombinase recT protein	809	16.07
ECs1936	hypothetical protein	233	21.46
ECs1944	hypothetical protein	788	15.23
ECs1956	hypothetical protein	599	15.03
ECs1957	hypothetical protein	290	17.24

<https://doi.org/10.1371/journal.pone.0202775.t001>

processing: DNA-invertase (ECs0284), transcriptional activator (ECs1588), and recombinase (ECs1933), restriction alleviation and modification enhancement protein (ECs1932), single-stranded DNA binding protein (ECs1587), and Icd-like protein (ECs1577).

The 4,383 genes without SNVs were also examined. Interestingly, we found that 2,002 (46%) of those genes were absent in the non-pathogenic *E. coli* K12 strain; in contrast, the corresponding proportion was lower among the SNV-containing genes (490 out of 1,201, 41%, *p* value = 0.001 by fisher exact test). These results suggested that the SNV-free genes of *E. coli* O157:H7 *E. coli* genome tended to be gained from sources other than the non-pathogenic *E. coli* genomes during its evolution.

Virulence gene profiling and virulent factors associated with the SNVs

Virulence factors (VFs) are essential in defining the pathogenicity of the *E. coli* O157:H7, which makes them the single most important gene family within the genome. Therefore, we profiled the shotgun sequencing data from our *E. coli* O157 strains against the 26K VFDB sequence records, using the ShortBRED pipeline. A total of 285 virulence genes were identified from our strain collection (S6 Table), and only 132 (46.3%, 132/285) of them were present in all of the profiled strains, indicating that genomes of the *E. coli* O157:H7 strains have dynamic and independently gained virulence genes. The common 132 virulence genes included *stx2B* and others that were characteristic to the *E. coli* O157:H7, such as the type III secretion system components (the Esp proteins), and the various flagellar proteins. One strain (KSU127) also contained the *stx1B* gene. Among all the identified virulence genes, 6 were only present in one of the profiled isolates (as listed in Table 2), including the *stx1B* (found in KSU127 isolate).

The *E. coli* O157:H7 strains were clustered based on the presence of the identified virulence genes. We found that strains from clades 6 and 7 tended to cluster well with the other strains from the same clade (Fig 5). However, strains from clade 8 spread out, except for one small cluster which contained only 5 strains.

We then examined the SNV status of the identified VF genes using our SNV annotation data. We first excluded the 51 VF genes that did not have the Sakai strain counterparts. Among the remaining ones, we found that 111 possessed no SNVs and 78 strains possessed low-density SNVs (with mean SNV density of 3 per 1,000q bases), suggesting that VF genes

Table 2. Virulence factors that were identified in single KSU isolate.

VFDB ID	Description	KSU isolate ID
VFG000836(gb NP_288672)	(stx1B) shiga-like toxin 1 subunit B encoded within prophage CP-933V [Stx (VF0206)] [<i>Escherichia coli</i> O157:H7 str. EDL933]	KSU127
VFG035574 (gi:218693685)	(aec27/clpV) ATP-dependent Clp proteinase Aec27 ATP-binding chain, with chaperone activity [ACE T6SS (CVF736)] [<i>Escherichia coli</i> 55989]	KSU263
VFG035974 (gi:15800037)	(ehaA) beta-barrel outer membrane protein [EhaA (SS109)] [<i>Escherichia coli</i> O157:H7 str. EDL933]	KSU055
VFG043048 (gi:16765293)	(ppdD) major pilin subunit [Hemorrhagic Coli pili (HCP) (AI102)] [<i>Escherichia coli</i> O157:H7 str. EDL933]	KSU126
VFG043263 (gi:16802741)	(flk) flagella biosynthesis regulator [peritrichous flagella (AI140)] [<i>Escherichia coli</i> O157:H7 str. EDL933]	KSU105
VFG034822 (gi:260843778)	(espR1)T3SS effector-like protein EspR-homolog[EspR1 (CVF696)]	KSU303

<https://doi.org/10.1371/journal.pone.0202775.t002>

belonged to a gene family with little genetic variances. We did observe three VF genes ECs0319 (hypothetical protein), ECs2665 (FliT), ECs1396 (AidA-I), with 12, 11, and 10 SNVs per 1,000 bases, respectively. The function of ECs0319 is still unknown, the ECs2665 (FliT) acts as transcriptional regulator and molecular chaperon to stabilize the flagella, and ECs1396 (AidA-I) facilitates adherence of the *E. coli* O157:H7 to the host cells. Their relatively higher variability might help the pathogen to adapt to different hosts.

Discussion

Direct link between the cattle-derived pathogenic *E. coli* O157:H7 and human infections, including severe complications, have been confirmed by many microbiological and epidemiological

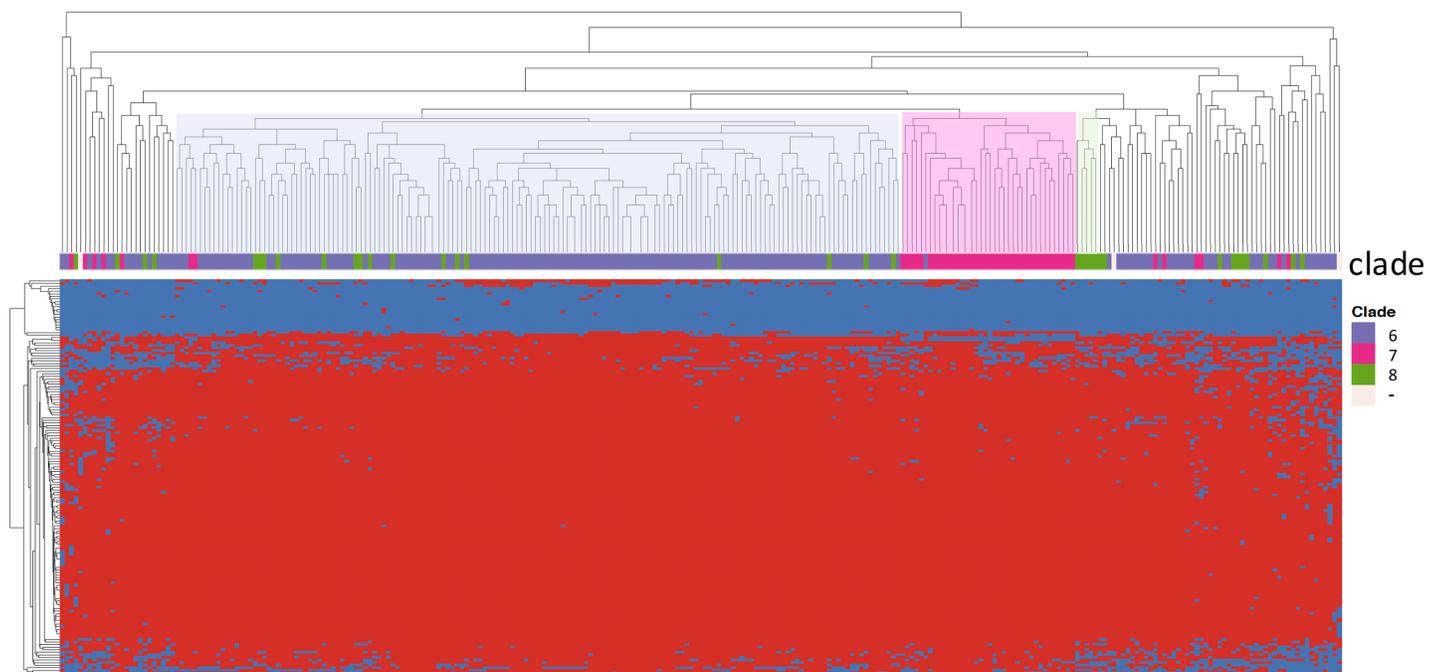


Fig 5. *E. coli* O157:H7 strains clustering based on the differences in the virulence gene profile. Clusters dominant with strains from the clades 6, 7, and 8 were highlighted in blue, pink and green, respectively.

<https://doi.org/10.1371/journal.pone.0202775.g005>

studies [48, 49]. The beef industry and the regulatory agencies have invested tremendous efforts to prevent entry of the *E. coli* O157:H7 into the food supply chain. A complicated issue about the *E. coli* O157:H7 control in cattle rearing practice and subsequent food safety management is its transient presence in asymptomatic cattle. In-depth evaluation of the genomic variability of such an important pathogen could provide insights into its pathogenic potentials and facilitate our understanding of the disease severity and search for new effective intervention solutions.

The goal of the present study was to better understand the genetic variations among the *E. coli* O157:H7 strains recovered from a single feedlot and immediate effects derived from those variants. We performed massive WGS on a total of 279 *E. coli* O157:H7 strains. With the high-quality, deep sequencing reads and available closed reference Sakai genome, we were able to profile all the SNVs across the whole *E. coli* O157:H7 genome region, which allowed us not only to construct the relatedness among all of the strains, but also locate the genes and the intergenic regions containing those genetic variances. Our results suggested that individual *E. coli* O157:H7 genome contains around 2,000 SNVs, which translate to one SNV per 2.8k bp within the genome.

The SNV-based phylogenomic tree confirmed the genetically homogeneous structure of *E. coli* O157:H7, and the extraordinary resolution of the tree basically clustered our *E. coli* O157:H7 collection into three major groups, essentially equivalent to the three defined clades [39]. We found that 12.5% of the collection was seen in clade 8 and 68.5% in clade 6, which have been reported to possess hyper-virulent phenotypes [50]. Particularly, the clade 8 is more likely to be associated with the severe disease cases. These results were quite different from those reported recently by our group based on a survey that involved 20 cattle farms from various regions in California, which revealed that the clade 8 was the most dominant group [51], probably reflecting certain geographic effects on the *E. coli* O157:H7 lineage distribution, as our current strains were from a feedlot located in the South Western region of the US. Two strains (KSU263/264) in our collection had distant phylogenetic relationship with all of the remaining ones and further analysis found that they were more closely clustered with the clinical O157 isolates 2011ELs. The three strains (KSU151, KSU155, and KSU247) that formed a subclonal group with clade 6 strains did not present typical clade-6 SNVs, suggesting that Manning clade classification method has its own limitations. However, it is beyond the scope of the current study to detail genome features with potential impact on their pathobiology and genetic classification.

Our SNV results did not reveal any core SNVs, and roughly 40% of the identified SNVs were actually absent from the reference Sakai genome, strongly suggesting the evolutionary plasticity within the *E. coli* O157:H7 clone and the high possibility of indel events during evolution. Nearly 80% of the SNVs were shown to have AF greater than 1%, indicating that the variant spectrum was stabilized. These results also provided additional evidence for the open pangenome model of the general *E. coli* species [52], which was inferred by gene content analysis, and highlighted the importance of *de novo* SNV calling during the evaluation of *E. coli* O157:H7 genome variations.

SNVs- enriched genomic windows along the Sakai genome were identified with our massive data. They were all within prophage loci. The *E. coli* O157:H7 Sakai genome contained 16 prophage loci with 11 of them intact [53]. Our data showed that that some of those loci were highly variable. For instance, prophage stx2_converting_II (genome region 1245903–1316397, about 70kb long) contained a total of 115 SNVs, and two genes ECs1242 (hypothetical protein) and ECs1232 (hypothetical protein) that were located at the 3' end of this prophage contributed to 36 SNVs. Similarly, the prophage entero_YYZ_2008 (genome region 1920485–1972367, about 52kb long) had a total of 170 SNVs, and its 5' end gene ECs1934 (exonuclease VIII RecE) alone contributed to 26 SNVs. Prophage loci contain the integrated bacteriophage

components, which have been proposed to play important roles in the emergence of the *E. coli* O157:H7 and their virulence [54]. The higher mutational rate of such loci suggested their frequent integration, which obviously would boost the transmission of the related virulence genes.

The detailed SNV annotation allowed us to examine individual genes for SNV density (SPK). We calculated gene-wise SPK and found that *E. coli* O157:H7 genes generally have less than 5 SPKs with peak at 1 SPK, suggesting a general functional stability of the basic bacterial genetics, although it won't rule out some type SNVs may drastically change the genome coding and the related biology. The small panel of highly variable genes contained many proteins with unknown functions. However, several proteins involved in DNA/RNA processing were revealed to contain high SPKs, strongly suggesting their roles in *E. coli* O157:H7 genome evolution and adaptation.

Clinical manifestations of the *E. coli* O157:H7 infections result from its virulence factors. Current VFDB collected about 26,000 VFs from various organisms. To get a comprehensive understanding about the virulence of the *E. coli* O157:H7 strains, we first generated a cluster-based secondary database using the VFDB peptide sequences which contained only the unique tags for any given VF gene or VF gene cluster. The secondary database allowed rapid screening of our massive WGS data against the large VFDB database records. Collectively, we found 285 VF genes among our 279 genomes. More than half of the identified VF genes ($n = 153$) were present in only a proportion of the strains, which might explain the wide disease spectrum of *E. coli* O157:H7 infections. The uneven presence of the VF genes in *E. coli* O157:H7 genomes also suggested that this gene family contributed significantly to the overall genomic variation among the strains. However, it is worth noting that there is a tradeoff between speed and resolution when the secondary database-based screening algorithm is applied. Some of the identified VF genes actually may represent several related VF genes.

Conclusions

Our analyses established SNV profiles for 279 *E. coli* O157:H7 strains collected from a single feedlot and the majority of the SNVs have never been reported before. Although geographically limited, we observed 3 out of 9 reported *E. coli* O157:H7 clades in our collection. Large collections and high-quality WGS data significantly expanded our knowledge about the genomic dynamics of this important pathogen, which may be subsequently applied to other pathogenic bacteria and more diverse sample collections. The data demonstrated the power of combing the genome profiling and gene functional analysis, and should prove useful in future framework development of epidemiological and biological studies of *E. coli* O157:H7.

Supporting information

S1 Fig. Histogram of the single nucleotide variant allele frequencies (AF) of *E. coli* O157:H7 strains isolated from cattle feces from a single feedlot in when Sakai genome was used as an annotation reference.

(TIF)

S2 Fig. Single nucleotide variant (SNV) allele frequencies and positions of *E. coli* O157:H7 strains isolated from cattle feces from a single feedlot when *E. coli* K12 strain was used as an annotation reference. (a) Histogram of the SNV allele frequencies (AF) when K12 was used as reference. (b) Number of SNVs per 2,000 bases along the Sakai strain genome. SNV-free region was zoomed in.

(TIF)

S1 Table. Number of SNVs in individual *E coli* K157:H7 isolates.
(XLSX)

S2 Table. Clade profile for the *E. coli* O157:H7 collected by KSU.
(XLSX)

S3 Table. Genomes related to the *E. coli* O157:H7 downloaded from the GenBank.
(XLSX)

S4 Table. SNVs identified from the KSU *E. coli* O157:H7 isolates using Sakai strain as reference.
(XLSX)

S5 Table. *E coli* O157:H7 genes that contain SNVs.
(XLSX)

S6 Table. Virulence factors identified from the KSU *E coli* O157:H7 isolates.
(XLSX)

S1 Text. Tree file for the *E. coli* O157 strains isolated from cattle feces from a single feedlot.
(TXT)

S2 Text. Tree file for the *E. coli* O157 strains isolated from cattle feces from a single feedlot and the relevant genomes downloaded from the GenBank.
(TXT)

Author Contributions

Conceptualization: Xin Gao, T. G. Nagaraja.

Data curation: Xin Gao, Xun Yang, Lance Noll, Xiaorong Shi.

Funding acquisition: Marc Allard, Eric Brown, T. G. Nagaraja, Jianghong Meng.

Investigation: Xin Gao, Xun Yang, Lance Noll, T. G. Nagaraja.

Methodology: Xin Gao, Xun Yang, Lance Noll, Xiaorong Shi, Jay Worley.

Supervision: Jianghong Meng.

Writing – original draft: Xin Gao, Xun Yang.

Writing – review & editing: Xin Gao, Lance Noll, Jay Worley, Marc Allard, Eric Brown, T. G. Nagaraja, Jianghong Meng.

References

1. Rangel JM, Sparling PH, Crowe C, Griffin PM, Swerdlow DL. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982–2002. *Emerg Infect Dis.* 2005; 11(4):603–9. <https://doi.org/10.3201/eid1104.040739> PMID: 15829201; PubMed Central PMCID: PMC3320345.
2. Zhou M, Guo Z, Duan Q, Hardwidge PR, Zhu G. *Escherichia coli* type III secretion system 2: a new kind of T3SS? *Vet Res.* 2014; 45:32. <https://doi.org/10.1186/1297-9716-45-32> PMID: 24641581; PubMed Central PMCID: PMC3977695.
3. Lim JY, Yoon J, Hovde CJ. A brief overview of *Escherichia coli* O157:H7 and its plasmid O157. *J Microbiol Biotechnol.* 2010; 20(1):5–14. PMID: 20134227; PubMed Central PMCID: PMC3645889.
4. Zhang D, Coronel-Aguilera CP, Romero PL, Perry L, Minocha U, Rosenfield C, et al. The Use of a Novel NanoLuc -Based Reporter Phage for the Detection of *Escherichia coli* O157:H7. *Sci Rep.* 2016; 6:33235. <https://doi.org/10.1038/srep33235> PMID: 27624517; PubMed Central PMCID: PMC5021930 Indiana which has licensed technology related to using PhiV10 and its derivatives

- for the detection of *E. coli* O157:H7. Authors Bruce Applegate, Dandan Zhang, Claudia P Coronel-Aguilera, Patricia L Romero, Lynda Perry, Udit Minocha and Carla Rosenfield are coinventors on recent disclosures of IP related to PhiV10 and its derivatives.
5. Teunis P, Takumi K, Shinagawa K. Dose response for infection by *Escherichia coli* O157:H7 from outbreak data. *Risk Anal.* 2004; 24(2):401–7. <https://doi.org/10.1111/j.0272-4332.2004.00441.x> PMID: 15078310.
 6. Riley LW, Remis RS, Helgeson SD, McGee HB, Wells JG, Davis BR, et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med.* 1983; 308(12):681–5. <https://doi.org/10.1056/NEJM198303243081203> PMID: 6338386.
 7. Wells JG, Davis BR, Wachsmuth IK, Riley LW, Remis RS, Sokolow R, et al. Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype. *J Clin Microbiol.* 1983; 18(3):512–20. PMID: 6355145; PubMed Central PMCID: PMCPMC270845.
 8. Greig JD, Ravel A. Analysis of foodborne outbreak data reported internationally for source attribution. *Int J Food Microbiol.* 2009; 130(2):77–87. <https://doi.org/10.1016/j.ijfoodmicro.2008.12.031> PMID: 19178974.
 9. Nataro JP, Kaper JB. Diarrheagenic *Escherichia coli*. *Clin Microbiol Rev.* 1998; 11(1):142–201. PMID: 9457432; PubMed Central PMCID: PMCPMC121379.
 10. Gyles CL. Shiga toxin-producing *Escherichia coli*: an overview. *J Anim Sci.* 2007; 85(13 Suppl):E45–62. <https://doi.org/10.2527/jas.2006-508> PMID: 17085726.
 11. Karmali MA. Infection by verocytotoxin-producing *Escherichia coli*. *Clin Microbiol Rev.* 1989; 2(1):15–38. PMID: 2644022; PubMed Central PMCID: PMCPMC358098.
 12. Cray WC Jr., Moon HW. Experimental infection of calves and adult cattle with *Escherichia coli* O157:H7. *Appl Environ Microbiol.* 1995; 61(4):1586–90. PMID: 7747972; PubMed Central PMCID: PMCPMC167413.
 13. Brown CA, Harmon BG, Zhao T, Doyle MP. Experimental *Escherichia coli* O157:H7 carriage in calves. *Appl Environ Microbiol.* 1997; 63(1):27–32. PMID: 8979335; PubMed Central PMCID: PMCPMC168298.
 14. LeJeune JT, Besser TE, Hancock DD. Cattle water troughs as reservoirs of *Escherichia coli* O157. *Appl Environ Microbiol.* 2001; 67(7):3053–7. <https://doi.org/10.1128/AEM.67.7.3053-3057.2001> PMID: 11425721; PubMed Central PMCID: PMCPMC92980.
 15. Avery LM, Killham K, Jones DL. Survival of *E. coli* O157:H7 in organic wastes destined for land application. *J Appl Microbiol.* 2005; 98(4):814–22. <https://doi.org/10.1111/j.1365-2672.2004.02524.x> PMID: 15752326.
 16. Reinstein S, Fox JT, Shi X, Alam MJ, Nagaraja TG. Prevalence of *Escherichia coli* O157:H7 in the American bison (*Bison bison*). *J Food Prot.* 2007; 70(11):2555–60. PMID: 18044434.
 17. Callaway TR, Edrington TS, Brabban AD, Keen JE, Anderson RC, Rossman ML, et al. Fecal prevalence of *Escherichia coli* O157, *Salmonella*, *Listeria*, and Bacteriophage Infecting *E. coli* O157:H7 in feedlot cattle in the Southern Plains region of the United States. *Foodborne Pathog Dis.* 2006; 3(3):234–44. <https://doi.org/10.1089/fpd.2006.3.234> PMID: 16972771.
 18. Sargeant JM, Amezcua MR, Rajic A, Waddell L. Pre-harvest interventions to reduce the shedding of *E. coli* O157 in the faeces of weaned domestic ruminants: a systematic review. *Zoonoses Public Health.* 2007; 54(6–7):260–77. <https://doi.org/10.1111/j.1863-2378.2007.01059.x> PMID: 17803515.
 19. Chui H, Chan M, Hernandez D, Chong P, McCorrister S, Robinson A, et al. Rapid, Sensitive, and Specific *Escherichia coli* H Antigen Typing by Matrix-Assisted Laser Desorption Ionization-Time of Flight-Based Peptide Mass Fingerprinting. *J Clin Microbiol.* 2015; 53(8):2480–5. <https://doi.org/10.1128/JCM.00593-15> PMID: 26019207; PubMed Central PMCID: PMCPMC4508425.
 20. Orskov I, Orskov F, Jann B, Jann K. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. *Bacteriol Rev.* 1977; 41(3):667–710. PMID: 334154; PubMed Central PMCID: PMCPMC414020.
 21. Manges AR, Harel J, Masson L, Edens TJ, Portt A, Reid-Smith RJ, et al. Multilocus sequence typing and virulence gene profiles associated with *Escherichia coli* from human and animal sources. *Foodborne Pathog Dis.* 2015; 12(4):302–10. <https://doi.org/10.1089/fpd.2014.1860> PMID: 25774654.
 22. Leung KT, Mackereth R, Tien YC, Topp E. A comparison of AFLP and ERIC-PCR analyses for discriminating *Escherichia coli* from cattle, pig and human sources. *FEMS Microbiol Ecol.* 2004; 47(1):111–9. [https://doi.org/10.1016/S0168-6496\(03\)00254-X](https://doi.org/10.1016/S0168-6496(03)00254-X) PMID: 19712352.
 23. Beutin L, Strauch A, Zimmermann S, Kaulfuss S, Schaudinn C, Mannel A, et al. Genetical and functional investigation of *fliC* genes encoding flagellar serotype H4 in wildtype strains of *Escherichia coli* and in a laboratory *E. coli* K-12 strain expressing flagellar antigen type H48. *BMC Microbiol.* 2005; 5:4. <https://doi.org/10.1186/1471-2180-5-4> PMID: 15663798; PubMed Central PMCID: PMCPMC548302.

24. Hirvonen JJ, Siitonen A, Kaukoranta SS. Usability and performance of CHROMagar STEC medium in detection of Shiga toxin-producing *Escherichia coli* strains. *J Clin Microbiol*. 2012; 50(11):3586–90. <https://doi.org/10.1128/JCM.01754-12> PMID: 22933601; PubMed Central PMCID: PMC3486272.
25. Laing C, Pegg C, Yawney D, Ziebell K, Steele M, Johnson R, et al. Rapid determination of *Escherichia coli* O157:H7 lineage types and molecular subtypes by using comparative genomic fingerprinting. *Appl Environ Microbiol*. 2008; 74(21):6606–15. <https://doi.org/10.1128/AEM.00985-08> PMID: 18791027; PubMed Central PMCID: PMC2576696.
26. Vernozzy-Rozand C, Montet MP, Bertin Y, Trably F, Girardeau JP, Martin C, et al. Serotyping, stx2 subtyping, and characterization of the locus of enterocyte effacement island of shiga toxin-producing *Escherichia coli* and *E. coli* O157:H7 strains isolated from the environment in France. *Appl Environ Microbiol*. 2004; 70(4):2556–9. <https://doi.org/10.1128/AEM.70.4.2556-2559.2004> PMID: 15066860; PubMed Central PMCID: PMC383133.
27. Sheng H, Lim JY, Knecht HJ, Li J, Hovde CJ. Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa. *Infect Immun*. 2006; 74(8):4685–93. <https://doi.org/10.1128/IAI.00406-06> PMID: 16861656; PubMed Central PMCID: PMC1539576.
28. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet*. 2012; 13(9):601–12. <https://doi.org/10.1038/nrg3226> PMID: 22868263; PubMed Central PMCID: PMC35049685.
29. Mayo B, Rachid CT, Alegria A, Leite AM, Peixoto RS, Delgado S. Impact of next generation sequencing techniques in food microbiology. *Curr Genomics*. 2014; 15(4):293–309. <https://doi.org/10.2174/1389202915666140616233211> PMID: 25132799; PubMed Central PMCID: PMC4133952.
30. Salipante SJ, SenGupta DJ, Cummings LA, Land TA, Hoogstraat DR, Cookson BT. Application of whole-genome sequencing for bacterial strain typing in molecular epidemiology. *J Clin Microbiol*. 2015; 53(4):1072–9. <https://doi.org/10.1128/JCM.03385-14> PMID: 25631811; PubMed Central PMCID: PMC4365209.
31. Rusconi B, Sanjar F, Koenig SS, Mammel MK, Tarr PI, Eppinger M. Whole Genome Sequencing for Genomics-Guided Investigations of *Escherichia coli* O157:H7 Outbreaks. *Front Microbiol*. 2016; 7:985. <https://doi.org/10.3389/fmicb.2016.00985> PMID: 27446025; PubMed Central PMCID: PMC4928038.
32. Ronholm J, Naseri N, Petronella N, Pagotto F. Navigating Microbiological Food Safety in the Era of Whole-Genome Sequencing. *Clin Microbiol Rev*. 2016; 29(4):837–57. <https://doi.org/10.1128/CMR.00056-16> PMID: 27559074; PubMed Central PMCID: PMC5010751.
33. Bai J, Shi X, Nagaraja TG. A multiplex PCR procedure for the detection of six major virulence genes in *Escherichia coli* O157:H7. *J Microbiol Methods*. 2010; 82(1):85–9. <https://doi.org/10.1016/j.mimet.2010.05.003> PMID: 20472005.
34. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *JComputBiol*. 2012; 19.
35. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014; 30(15):2114–20. <https://doi.org/10.1093/bioinformatics/btu170> PMID: 24695404; PubMed Central PMCID: PMC4103590.
36. Ambroise J, Bearzatto B, Robert A, Govaerts B, Macq B, Gala JL. Impact of the spotted microarray pre-processing method on fold-change compression and variance stability. *BMC Bioinformatics*. 2011; 12:413. <https://doi.org/10.1186/1471-2105-12-413> PMID: 22026942; PubMed Central PMCID: PMC3224096.
37. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*. 2015; 31(17):2877–8. <https://doi.org/10.1093/bioinformatics/btv271> PMID: 25913206.
38. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012; 6(2):80–92. <https://doi.org/10.4161/fly.19695> PMID: 22728672; PubMed Central PMCID: PMC3679285.
39. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, Ouellette LM, et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci U S A*. 2008; 105(12):4868–73. <https://doi.org/10.1073/pnas.0710834105> PMID: 18332430; PubMed Central PMCID: PMC2290780.
40. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res*. 2002; 12(4):656–64. <https://doi.org/10.1101/gr.229202> PMID: 11932250; PubMed Central PMCID: PMC187518.
41. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. *PLoS Comput Biol*. 2015; 11(12):

- e1004557. <https://doi.org/10.1371/journal.pcbi.1004557> PMID: 26682918; PubMed Central PMCID: PMC4684307.
42. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res.* 2016; 44(D1):D694–7. <https://doi.org/10.1093/nar/gkv1239> PMID: 26578559; PubMed Central PMCID: PMC4702877.
 43. Trees E, Strockbine N, Changayil S, Ranganathan S, Zhao K, Weil R, et al. Genome Sequences of 228 Shiga Toxin-Producing *Escherichia coli* Isolates and 12 Isolates Representing Other Diarrheagenic *E. coli* Pathotypes. *Genome Announc.* 2014; 2(4). <https://doi.org/10.1128/genomeA.00718-14> PMID: 25103754; PubMed Central PMCID: PMC4125765.
 44. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl Acad Sci U S A.* 2011; 108(50):20142–7. <https://doi.org/10.1073/pnas.1107176108> PMID: 22135463; PubMed Central PMCID: PMC3250189.
 45. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. Genome signatures of *Escherichia coli* O157:H7 isolates from the bovine host reservoir. *Appl Environ Microbiol.* 2011; 77(9):2916–25. <https://doi.org/10.1128/AEM.02554-10> PMID: 21421787; PubMed Central PMCID: PMC3126413.
 46. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis.* 2015; 61(3):305–12. <https://doi.org/10.1093/cid/civ318> PMID: 25888672; PubMed Central PMCID: PMC4542925.
 47. Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, et al. Utility of Whole-Genome Sequencing of *Escherichia coli* O157 for Outbreak Detection and Epidemiological Surveillance. *J Clin Microbiol.* 2015; 53(11):3565–73. <https://doi.org/10.1128/JCM.01066-15> PMID: 26354815; PubMed Central PMCID: PMC4609728.
 48. Yang SC, Lin CH, Aljuffali IA, Fang JY. Current pathogenic *Escherichia coli* foodborne outbreak cases and therapy development. *Arch Microbiol.* 2017; 199(6):811–25. <https://doi.org/10.1007/s00203-017-1393-y> PMID: 28597303.
 49. Paton JC, Paton AW. Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. *Clin Microbiol Rev.* 1998; 11(3):450–79. PMID: 9665978; PubMed Central PMCID: PMC488891.
 50. Amigo N, Mercado E, Bentancor A, Singh P, Vilte D, Gerhardt E, et al. Clade 8 and Clade 6 Strains of *Escherichia coli* O157:H7 from Cattle in Argentina have Hypervirulent-Like Phenotypes. *PLoS One.* 2015; 10(6):e0127710. <https://doi.org/10.1371/journal.pone.0127710> PMID: 26030198; PubMed Central PMCID: PMC4452545.
 51. Worley JN, Flores KA, Yang X, Chase JA, Cao G, Tang S, et al. Prevalence and Genomic Characterization of *Escherichia coli* O157:H7 in Cow-Calf Herds throughout California. *Appl Environ Microbiol.* 2017; 83(16). <https://doi.org/10.1128/AEM.00734-17> PMID: 28550057; PubMed Central PMCID: PMC5541215.
 52. 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; PubMed Central PMCID: PMC2566221.
 53. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. *Nucleic Acids Res.* 2011; 39(Web Server issue):W347–52. <https://doi.org/10.1093/nar/gkr485> PMID: 21672955; PubMed Central PMCID: PMC3125810.
 54. Ohnishi M, Kurokawa K, Hayashi T. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? *Trends Microbiol.* 2001; 9(10):481–5. PMID: 11597449.