Prophage-dependent recombination drives genome structural variation and phenotypic heterogeneity in *Escherichia coli* O157:H7

Stephen F. Fitzgerald[1]*, Nadejda Lupolova[1], Sharif Shaaban[1], Timothy J. Dallman[2], David Greig[2], Lesley Allison[3], Sue C. Tongue[4], Judith Evans[4], Madeleine K. Henry[4], Tom N. McNeilly[5], James L. Bono[6] and David L. Gally[1]*.

[1] Division of Immunity and Infection, The Roslin Institute and R(D)SVS, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK.

[2] Gastrointestinal Bacterial Reference Unit, 61 Colindale Avenue, Public Health England, NW9 5EQ London, UK.

[3] Scottish E. coli O157/VTEC Reference Laboratory, Department of Laboratory Medicine, Royal Infirmary of Edinburgh, 51 Little France Crescent, Edinburgh EH16 4SA, UK.

[4] Epidemiology Research Unit (Inverness), Department of Veterinary and Animal Science, Northern Faculty, Scotland's Rural College (SRUC), Scotland, IV2 5NA, UK.

[5] Moredun Research Institute, Pentlands Science Park, Bush Loan, Penicuik, EH26 OPZ, UK.

[6] United States Department of Agriculture, Agricultural Research Service, US Meat Animal Research Center, Clay Center, Nebraska, USA.

Running Title: Large genome rearrangements in *E. coli* O157

Keywords: genome structure; duplication; inversion; Shiga toxin; type 3 secretion; PFGE; optical mapping; *E. coli* O157; cattle; prophage

*Corresponding author: Prof. David L. Gally, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh EH25 9RG, UK. Telephone: +44 (0)131 6519242; email: dgally@ed.ac.uk

Abstract

The human zoonotic pathogen *Escherichia coli* O157 is defined by its extensive prophage repertoire including those that encode Shiga toxin, the factor responsible for inducing life-threatening pathology in humans. As well as introducing genes that can contribute to the virulence of a strain, prophage can enable the generation of large-chromosomal rearrangements (LCRs) by homologous recombination. This work examines the types and frequencies of LCRs across the major lineages of the O157 serogroup and defines the phenotypic consequences of specific structural variants. We demonstrate that LCRs are a major source of genomic variation across all lineages of *E. coli* O157 and by using both optical mapping and ONT long-read sequencing demonstrate that LCRs are generated in laboratory cultures started from a single colony and particular variants are selected during animal colonisation. LCRs are biased towards the terminus region of the genome and are bounded by specific prophages that share large regions of sequence homology associated with the recombinational activity. RNA transcriptional profiling and phenotyping of specific structural variants indicated that important virulence phenotypes such as Shiga toxin production, type 3 secretion and motility are affected by LCRs. In summary, *E. coli* O157 has acquired multiple prophage regions over time that act as genome engineers to continually produce structural variants of the genome. This structural variation is a form of epigenetic regulation that generates sub-population phenotypic heterogeneity with important implications for bacterial adaptation and survival.

Author Summary

Escherichia coli has an 'open genome' and has acquired genetic information over evolutionary time, often in the form of bacteriophages that integrate into the bacterial genome (prophages). *E. coli* O157 is a clonal serogroup that is found primarily in ruminants such as cattle but can cause life-threatening infections in humans. *E. coli* O157 isolates contain multiple prophages including those that encode Shiga-like toxins which are responsible for the more serious disease associated with human infections. We show in this study that many of these prophages exhibit large regions of sequence similarity that allow rearrangements to occur in the genome generating structural variants. These occur routinely during bacterial culture in the laboratory and the variants are detected during animal colonization. The variants generated can give the bacteria altered phenotypes, such as increased motility or toxin production which can be selected in specific environments and therefore represent a highly dynamic mechanism to generate variation in bacterial populations without a change in overall gene content.

Introduction

Bacterial viruses, termed prophage, that incorporate their genomes onto the bacterial chromosome are major drivers of bacterial genome evolution, host and niche adaptation and virulence [1-3]. Prophage integration directly benefits the bacterial host by conferring resistance against other lytic viruses [4], by carriage of virulence factors, including toxins and effector proteins [1, 5], enzymes involved in stress resistance [6] and the expression both gene regulators and sRNAs capable of influencing the host gene regulatory network [2, 7]. Here we examine the impact prophages have on the structure of the bacterial genome through the generation of large-chromosomal rearrangements (LCRs).

Escherichia coli O157:H7 is a significant human zoonotic pathogen originating from ruminant hosts, especially cattle [8]. Over evolutionary time, numerous prophage (typically 16 – 25) have integrated into the genomes of *E. coli* O157 strains with an integration bias towards the terminus (Ter) of replication [9]. Acquisition of these prophage, many of which are closely related λ-like phage, has driven the evolution of this pathogen by carriage of virulence genes including secreted effector proteins, sRNAs involved in virulence gene regulation and [7, 10], importantly, these prophage include those that encode Shiga toxin (Stx) subtypes. Stx toxins are the main mediators of vascular endothelial cell killing in infected humans [11] and the resulting damage can lead to haemolytic uremic syndrome (HUS), often fatal, or lead to life-long kidney and brain damage [12-14]. *E. coli* O157:H7 strains are divided into three phylogenetically distinct lineages (I, I/II and II) with those that represent a serious threat to human health belonging to lineage I or Lineage I/II and the majority encode two sub-types of Stx, Stx2a and Stx2c. Stx2a is generally associated with more serious disease [11, 15-18] and the emergence of *E. coli* O157:H7 as a zoonotic threat correlates with the introduction of Stx2a-encoding prophage into the *E. coli* O157 cattle population approximately 50 years ago [11].

There is published evidence that *E. coli* O157 type strain EDL933 can undergo large-chromosomal rearrangements (LCRs), mainly inversions [19, 20], with these rearrangements being flanked by prophages. LCRs, such as inversions, duplications and translocations, occur by homologous recombination between repeat sequences on the same chromosome [21]. While LCRs arising between ribosomal *rrn* operons, pathogenicity islands and insertion sequence (IS) elements have been associated with speciation, diversification, outbreaks and immune evasion in bacteria [1, 22] few studies have examined LCRs arising from inter-prophage recombination and their impact on phenotype.

In this study we demonstrate that prophage-mediated LCRs are a major source of genomic variation across all lineages of *Escherichia coli* O157. We show that alternate chromosomal conformations are generated during laboratory culture and are selected during host colonisation. Specific LCRs were associated with changes in virulence phenotypes and we therefore propose that the generation of LCRs within *E. coli* O157 populations *in vivo* facilitates phenotypic heterogeneity and niche

adaptation, include host colonisation. Prophage act as genome engineers by driving conservative rearrangements leading to sub-populations with distinct phenotypes that can provide an advantage in different environments.

Results

LCRs shape *E. coli* O157 genome evolution

To examine the extent of genomic diversity generated by LCRs in the *E. coli* O157 clonal group, we examined the whole genome sequences of 72 isolates, the majority of which were generated by PacBio long-read sequencing (Table S1). Strains analysed were representative of the main *E. coli* O157 lineages (I, I/II and II) and included multiple sub-Lineage Ic, PT21/28 isolates which have been responsible for the majority of serious human infections in the UK over the last two decades [11]. This genome dataset included previously sequenced complete genomes from each lineage, including strains Sakai (NC_002695.2), EDL933 (CP008957.1) and TW14359 (CP001368) (Table S1).

Pairwise alignment of all 72 genomes identified LCRs, predominantly large inversions, as a common source of genomic variation between isolates within each *E. coli* O157 lineage with the exception of lineage I/II (Supplementary Figure S1). In addition, each genome was individually aligned against a representative reference strain from each of four lineages and the chromosomal loci of all LCRs > 50 kb were mapped (Fig 1A-D). The reference strains were: Strain 9000 (Lineage 1c), Sakai (Lineage 1a), TW14359 (Lineage I/II) and Strain 180 (Lineage II).

LCRs > 50 kb were frequently identified in lineages Ia, Ic and II irrespective of the reference strain used for alignment, however it was evident that Lineage I/II strains exhibited less variation (Figure 1). Strains from Lineage 1c and Lineage II exhibited the most variation at this macro level with an average of 43 and 37 LCRs identified, respectively, (Table. 1 and Fig. 1). Strains from Lineage 1a were less variable with an average of 14 LCRs identified across all strains and the least genomic variation with respect to the reference strains was observed for strains from Lineage I/II with an average of just 2.5 LCRs identified in a single strain, F8492. We note that strain F8492 was a singleton isolate that grouped closely with our other representative Lineage I/II strains (Supplementary Figure S2). Lineage I/II strains were also the least variable when the number of LCRs identified were corrected to account for the unequal number of strains analysed within each lineage (Table. 1). To examine this further, we plotted the average size of all LCRs with a lower cut-off of >20 kb that could be detected in each strain relative to the four reference genomes (Supplementary Figure S2 A – D). At this lower cut-off, LCRs ranging between 20 kb and 30 kb were identified in Lineage I/II that were generally consistent across all Lineage I/II strains relative to each reference genome. These results indicate that the macro genome conformation of Lineage I/II strains

is highly conserved. While LCRs can occur within Lineage I/II strains they have a reduced capacity to generate larger LCRs > 30 kb compared with the two other lineages.

For all lineages, LCRs were biased toward the chromosomal terminus of replication (Ter) with the majority located between 2 Mbp – 3.5 Mbp (Fig. 1). The largest LCR identified was a 1. 4 Mbp inversion which was detected in Lineage Ic strain Z1615 (Fig. 1 and Supplementary Figure. S1). The average length of LCRs detected ranged between 109 – 376 Kbp depending on which reference strain was used for alignment with the largest LCRs detected within lineage Ic strains when aligned against lineage Ic reference strain 9000 (Table. 1 and Supplementary Figure. S3A). Mapping the chromosomal position of prophages within each reference genome further demonstrated that most LCRs were bounded by prophages (marked in red in comparison strain, Fig. 1). Furthermore, many of the LCRs identified had prophage Stx2c (ΦStx2c) as a boundary, particularly those occurring within Lineage Ic strains.

Table 1. Mean number and size of LCRs relative to each reference genome

	Total No. of LCRs			
Reference	**Lineage Ia**	**Lineage Ic**	**Lineage I/II**	**Lineage II**
9000	18	36	2	43
Sakai	15	47	2	38
TW14359	13	47	2	34
180	13	43	4	41
Mean	14.75	43.25	2.5	39
	LCRs per lineage/strain			
9000	1.64	1.16	0.2	2.15
Sakai	1.36	1.52	0.2	1.9
TW14359	1.18	1.52	0.2	1.7
180	1.18	1.39	0.4	2.05
Mean	1.34	1.40	0.25	1.95
	Average LCR length (bp)			
9000	109705	376229	126954	110504
Sakai	151151	143497	142760	114507
TW14359	155589	134568	149038	127925
180	131237	144404	119660	128622
Mean	136920	199675	134603	120390

LCRs map to repeated regions of homology on prophage

Mechanistically, chromosomal inversions typically involve recombination between inverted repeat regions of homologous sequences [22, 23]. As inversions were the dominant LCR identified in our analysis (Fig 1), we mapped the chromosomal position and direction of all homologous regions for each *E. coli* O157 strain (Fig. 2 and Table. S2). To avoid detection of the numerous IS elements present in *E. coli* O157 genomes [24] we restricted our analysis to regions that shared ≥ 98 % sequence homology, were ≥ 5000 bp and occurred in the chromosome with a frequency ≥ 2. Repeat regions were unequally distributed throughout the chromosome with a bias toward Ter and were

conserved as inverted repeats at either side of Ter (Fig 2A). When each genome was subdivided into 1 Mbp domains, significantly more repeats were located within the 2 – 3 Mbp domain ($p < 0.0001$) adjacent to Ter than any other domain of the chromosome (Supplementary Figure. S3B). Significantly more repeats were also located within the 3 – 4 Mbp domain ($p < 0.0001$) adjacent to Ter than the 1 – 2 Mbp, 4 – 5 Mbp and 5 – 6 Mbp regions but not the 0 – 1 Mbp domain ($p = 0.71$). All repeat regions identified in the terminal half of the chromosome mapped within prophage (Fig 2B and Fig 2C) and specific combinations of these repeated regions matched the boundaries for identified LCRs. For example, specific recombination between regions 1a and 1b of Strain 9000 in Fig 2B would generate the LCR present in isogenic strain Z1767 and recombination between 2a and 2b would generate the LCR present in isogenic strain Z1615. These results indicate that homologous prophage sequences are hotspots for recombination resulting in the generation LCRs in *E. coli* O157 strains.

It was evident that specific combinations of inverted repeat regions were present in the different lineages and sub-lineages of *E. coli* O157 (Fig 2A). We reasoned that the frequency of recombinational events would be greater in strains with more homologous repeat regions and *vice versa*. Indeed, strains from Lineage Ic, in which the greatest number of LCRs were identified (Table. 1), had significantly more repeat regions > 5000 bp ($p < 0.05$) (Supplementary Figure S4A) and > 8000 bp ($p < 0.0001$) (Supplementary Figure S4B) than those from any other lineage. Conversely, Lineage I/II strains, in which only a single LCR was identified, had fewer homologous repeat regions ≥5000 bp than strains from any other lineage (Supplementary Figure S4A) and significantly less homologous repeat regions ≥ 8000 bp ($p < 0.01$) (Supplementary Figure S4B).

LCRs underpin PFGE type expansion in Lineage Ic PT21/28 strains

In the United Kingdom, PT21/28 strains from Lineage Ic have arisen as the dominant PT associated with severe human infections over the last 20 years [11]. Based on standard pulsed-field gel electrophoresis (PFGE) typing methods, PT21/28 isolates have expanded from an initial 5 PFGE types (Profiles A - E, personal communication from Dr Lesley Allison Scottish *E. coli* reference laboratory-SERL) present in the UK in 1994 to >30 distinct PFGE profiles (Fig. 3A and Supplementary Figure. S5) by 2013 when PFGE was replaced by MLVA analysis. LCRs were shown to generate changes in the PFGE type of strain EDL933 [19], we therefore determined if LCRs also underpinned the PFGE type expansion seen in PT21/28 strains. We sequenced ten PT21/28 isolates by PacBio long-read sequencing that differed in PFGE type. Strains were selected from throughout the PT21/28 core SNP based phylogeny (Supplementary Figure. S5) and the dataset included two isolates with identical SNP addresses (Z910 and Z563; zero SNP differences in the core genome) but with distinct PFGE profiles.

Sequence analysis showed that all ten strains differed by < 70 SNPs in their core genomes (Supplementary Figure. S5). Although examples of phage gain/loss (n = 2) were apparent, pairwise

whole genome comparisons showed that LCRs were the dominant source of genomic variation at the macro scale (Fig. 3B). Reference laboratories specializing in STEC diagnostics in the UK used AvrII and/or XbaI restriction enzymes when determining the PFGE type for an isolate. When all AvrII restriction sites were mapped in each isolate (Fig. 3B) it was evident that the loci of most sites were strongly conserved. However significant strain variation in AvrII loci was observed within the Ter region of the chromosome that was associated with LCRs. For example, strains Z910 and Z563, which were identical at the core SNP level, differed by a single 1.2 Mbp chromosomal inversion that involved recombination with ΦStx2c and resulted in the repositioning of four AvrII sites. Additional sequences containing AvrII sites present in some strains but not in others were identified (Fig. 3B) however these were rare. The majority of AvrII loci variation and therefore PFGE type variation was generated by LCRs.

To confirm that the variation in AvrII loci generated by LCRs observed in our PacBio assemblies matched the actual chromosome configuration of each isolate we determined the PFGE profile for each strain after AvrII restriction digestion (Fig. 3C) and compared it to *in silico* AvrII digests of their respective Pac-Bio assemblies (Fig. 3D). Both *in vivo* and *in silico* AvrII digestion patterns were matched for 9/10 strains analysed confirming the presence of those LCRs identified by PacBio long-read sequencing and the rearrangement of AvrII loci by these LCRs to generate different PFGE types. The exception was strain Z892 in which an unexplained digestion product was present after *in vivo* digestion that was not predicted from the PacBio sequence.

Based on these results we propose that the majority of the PFGE variation amongst PT21/28 strains, as depicted in Fig. 3A and Supplementary Figure. S5, is generated by LCRs. It was also evident from the PFGE analyses that the strains cultured under these laboratory conditions had the majority of their genomes in a single confirmation as there was no evidence of weak secondary bands in the gel restriction patterns (Fig. 3C). Of note, the most frequently occurring PT21/28 strain PFGE profile was type 'C' later defined as profile A_11b (Fig. 3A and Fig. S5). Phylogenetically, this specific profile re-occurs throughout the sub-lineage indicating that it is likely an ancestral confirmation or strains can repeatedly return to this chromosome conformation.

***In vivo* occurrence of LCRs during host colonisation**

Previously, we carried out a series of published and unpublished *in vivo* cattle colonization studies focused on *E. coli* O157 strain 9000 [25, 26]. To determine if LCRs are present during animal colonization we compared isolates collected from two separate colonization studies by PacBio long-read sequencing and AvrII PFGE profiling. This isolate set were all derivatives of the original wildtype strain 9000 and included inoculum and recovered isolates (Supplementary Table S1).

Pairwise whole genome comparisons of strains 9000 and Z1615 from Trial 1 (Fig. 4A) showed a 1.4 Mbp inversion had occurred in derivative strain Z1615 relative to strain 9000. As outlined in Fig. 2B

the boundaries of this LCR mapped to large inverted repeat sequences within prophage located either side of Ter. Distinct PFGE profiles were observed for strains 9000 and Z1615 following AvrII digestion, each matched their respective *in silico* AvrII digestion profiles (Supplementary Figure S6A) and confirmed the presence of the LCR identified in Z1615. No evidence of secondary bands diagnostic of Z1615 chromosomal conformation in the PFGE profile of strain 9000 were observed indicating this LCR occurred or was selected during colonization to generate strain Z1615. To determine how frequently this LCR occurs we analysed a further eleven recovered isolates from two experimental trials (Trial 1 and Trial 3) in which strain 9000 was the inoculum by PFGE (Supplementary Figure S6B). Isolates were collected from a number of different animals and dates (Supplementary Table S1). Three additional isolates of the 11 tested matched the PFGE profile of Z1615 indicating an *in vivo* selection for this LCR in the bovine host.

Two additional LCRs were identified from the five isolates examined from Trial 2 (Fig. 4B) both of which involved recombination with the Stx2c prophage (ΦStx2c). A 220 kbp inverted duplication was identified in strain Z1723. The duplicated region was flanked by repeat sequences from within prophage located at 2.2 Mbp and 2.4 Mbp (Supplementary Table S2) relative to OriC and inserted into ΦStx2c (3.4 Mbp) bisecting the Stx2c prophage (Fig. 4B and Fig. 4C). A second 1.2 Mbp inversion was identified in strain Z1767 that also involved recombination between repeat sequences within the same prophage located at 2.2 Mbp and ΦStx2c (Fig. 4B and Fig. 4C). PFGE analysis confirmed the presence of the LCRs in Z1723 and Z1767 (Supplementary Figure. S6A).

Real-time occurrence of LCRs during *in vitro* laboratory culture

We investigated if LCRs could be generated and detected in real-time following standard laboratory culture of bacteria in LB media. To increase the sensitivity of detection we applied both Oxford Nanopore Technologies (ONT) long-read sequencing and optical mapping to detect LCRs in strains from animal colonization Trials 1 and 2.

The wildtype parental strain 9000 was first sequenced using ONT and searched for reads that aligned to the LCRs identified in variant strains Z1615, Z1767 or Z1723. Aligning strain 9000 reads to the Z1615 genome, a total of 5 reads were found that matched the identified 1.4 Mb inversion boundary at 1.95 Mb relative to OriC and a single read that matched the inversion boundary at 3.35 Mb. These reads were abundant at approximately 2 % and 0.33 %, respectively, of the total reads across the same region that mapped directly to strain 9000. Similarly aligning 9000 reads to the Z1767 genome, a single read (0.4 % abundance) was found that matched the 1.2 Mb inversion boundary at 2.25 Mb relative to OriC and three reads (1.2 % abundance) that matched the inversion boundary within ΦStx2c at 3.45 Mb. No reads were found that mapped to the 220 kb duplication in Z1723.

Next we analysed strains Z1723 and Z1767 using Bionano Irys optical mapping (Figure 5 and Supplementary Figure S6) to identify additional LCRs that occur during growth in LB medium. Cultures of each strain were started from single colonies and chromosomes were extracted during late exponential phase cultures (OD600 = 0.7). Structural variant (SV) analysis was performed to detect all novel genome restriction maps within the cultured populations of Z1723 and Z1767 that did not map directly to an *in silico* generated map of the parental strain 9000 reference genome (Fig. 5).

Optical mapping showed that both strains had mixed population structures when cultured *in vitro*. SV analysis confirmed the same 220 kb inverted duplication was present in the Z1723 population that was identified by PacBio sequencing and PFGE (Fig. 5A). This hybrid structural variant mapped 5' – 3' between 2.24 – 2.46 Mb and 3' – 5' between 3.26 – 3.46 Mb to Strain 9000 further confirming the presence of the inverted duplication within ΦStx2c at 3.4 Mb. A 1.2 Mbp inversion relative to strain 9000 (Fig. 5B) was also identified in Z1723. This inversion matched the 1.2 Mbp inversion seen in strain Z1767 (Fig. 4B) with boundaries in prophage located at 2.2 Mbp and 3.4 Mbp (ΦStx2c). PFGE analysis of two separate Z1723 freezer stocks (Supplementary Figure S6A) shows that the 220 kbp inverted duplication is the dominant genome conformation present with no evidence of secondary bands indicative of the Z1767 inversion. We therefore assume that the 1.2 Mbp inversion detected in the Z1723 population by optical mapping is a minority population below the limit of detection by PFGE.

SV analysis of Z1767 identified the expected 1.2 Mbp inversion relative to strain 9000 (Supplementary Figure S7A) as determined from Pac-Bio sequencing and identified a novel 140.5 kbp inverted duplication within the cultured population (Supplementary Figure S7B). The duplicated region spanned 2.1 – 2.24 Mbp relative to OriC and was flanked by prophage sequence (2.2 Mbp) and an IS66 sequence located within the O-Island 48 [27]. This duplicated region also inserted in an inverted orientation within the Stx2c prophage further highlighting ΦStx2c as a hotspot for recombinational events leading to LCRs.

Changes in bacterial gene expression and phenotypes associated with LCRs

Using the structural variants of strain 9000 (Z1723, Z1767, Z1615) generated during *in vivo* colonization we examined if the identified LCRs impacted strain phenotypes. The global transcriptomes of strain 9000 and each structural variant strain (Z1723, Z1767, Z1615) were first compared by RNAseq for two growth conditions: nutrient rich LB medium and minimal M9 medium. PCA analysis showed there was little discernible difference between the transcriptomes of each strain when cultured in LB (Supplementary Figure S8A) however the transcriptome of strain Z1723, containing a 220kbp inverted duplication, was distinct from strain 9000 and the other variants in M9 (Supplementary Figure S8B). Differential changes in gene expression were modest (Supplementary Table S3) although a gene dosage effect was apparent across the region of duplication with an

increase in expression observed for 66 of the duplicated genes when mapped to the genome of WT strain 9000 (Fig. 6A). There was also a marked effect on the expression of genes within the Stx2a prophage (ΦStx2a) rather than the Stx2c prophage (ΦStx2c) into which the 220 kbp duplication had inserted (Fig. 6A).

As Stx2 toxin is the primary virulence factor of *E. coli* O157 strains leading to HUS, we tested if the observed differential transcription within ΦStx2a in Z1723 affected Stx2a expression, production and activity compared with other structural variants. For each phenotype Z1723 was compared with Trial 2 variants Z1766 and Z1767. Strains 9000 and Z1615 were excluded due to the previously documented [25] inactivation of the *stx2a* gene by an IS element, IS629. Expression of *stx2a* was increased in Z1723 compared to both Z1766 and Z1767 (Fig. 6B) and this manifested as a significant increase in total Stx2 toxin (Fig. 6C) and cytotoxic killing of Stx2 susceptible Vero cells (Fig. 6D).

We have previously shown that lysogeny with Stx2 prophages negatively regulates the LEE type III secretion system (T3S) [28] and demonstrated that a large duplication may have influenced the fitness of two closely related outbreak strains [29]. As the 220 kbp duplication in Z1723 interrupted ΦStx2c and increased expression of ΦStx2a genes we examined T3S and assessed the competitive fitness for strains Z1723, Z1766 and Z1767 (Fig. 7). Transcriptional *gfp* fusions to the LEE master regulator, *ler*, and LEE4 encoded *sepL* were introduced into each strain and expression was monitored in MEM-HEPES medium (OD600 = 0.8). Expression of both *ler* and *sepL* was decreased in Z1723 compared to Z1766 and Z1767 (Fig. 7A). There was also marked difference in the levels of the T3S secreted protein, EspD, which could not be detected in the culture supernatant of Z1723 (Fig. 7B).

The competitive fitness of strains from Trial 1 (9000 and Z1615) and Trial 2 (Z1723, 1766, 1767) was assessed by paired co-culturing in M9 media. In M9 media Z1723 significantly outcompeted the structural variants Z1766 and Z1767 as mean fitness indices (f.i.) of 0.89 and 0.93 were recorded, respectively (Fig. 7C) compared with control, f.i. = 1. No significant difference in fitness was observed between trial 1 strains in M9 (Fig. 7C). Finally, we measured the motility of strains 9000 and each structural variant on tryptone swarm plates (Fig. 7D). For strains isolated from calf trial 2 no difference in motility between Z1723 and Z1767 was observed however Z1766 was significantly more motile than both variants. Z1615 from calf trial 1 was also significantly more motile than WT strain 9000. These data provide evidence that LCRs can impact important *E. coli* O157 phenotypes involved in host colonisation and disease.

Discussion

Phenotypic heterogeneity within isogenic populations of microorganisms is used as a 'bet-hedging' survival strategy to cope with sudden fluctuations in environmental conditions and can lead to a division of labour between individuals that raises group fitness [30-33]. We have demonstrated that *E. coli* O157 can generate such heterogeneity through LCRs occurring between homologous

prophage sequence *in vivo* and *in vitro*. As originally demonstrated for *E. coli* O157 strain EDL933 [19] the LCRs we have now documented across the serogroup are bounded by specific prophages clustered towards the terminus of the genome. Chromosomal inversions involving the Ter region that lead to replichore imbalance can stall or stop replication forks and induce SOS [34]. Due to the spatial distribution of prophages involved in LCRs, the main large inversions we have identified do not generate major changes in replichore size. However even minor changes could impact growth rate and phenotypes as seen with the LCR specific phenotypes identified in this study affecting virulence gene expression, fitness and motility.

Large prophage homologous repeats (> 5000 bp) were identified at the boundaries of LCRs which provide ample sequence substrate for recombination. In addition to RecABCD-mediated recombination, *E. coli* O157 strains also carry multiple λ-like phage, including Stx phage, many of which encode their own Rad52-like recombinase enzymes such as Redβ [35-37]. Whether the formation of LCRs in *E. coli* O157 strains is host or phage mediated is unknown. Irrespective of the recombination system involved, the generation of LCRs would require a double-strand break (DSB) in one or more of the phage at their boundaries. It is interesting to speculate that double-strand breaks (DSBs) within phage are a primary driver of recombinational repair in bacteria via the SOS response that is also required for prophage-based expression of Stx. Strains of *E. coli* O157 PT21/28 constitutively express Stx2 [25] and therefore the rate of occurrence of DSBs, RecA-mediated Stx expression and LCR formation may be interconnected.

The Stx2c-encoding prophage was shown to be present across the different *E. coli* O157 lineages without much variation compared to Stx2a encoding prophages [9, 11]. In the present study it is a primary architect of many of the LCRs and as such may be subject to positive selection. One structural variant of PT21/28 strain 9000 was a duplication from one side of the chromosome inserted into the Stx2c terminase gene region on the other side of the genome. This was of particular interest as this large region of duplicated homology would stimulate inversions and also recombination resolving back to the original confirmation. A similar duplication has been sequenced in two closely related strains associated with sequential *E. coli* O157 outbreaks at a single restaurant [19]

The ONT long-read sequencing and optical mapping results provide evidence that LCRs are continuously generated at very low levels. The estimation from the ONT long-read sequencing of strain 9000 was between 1 – 2 % of the population when cultured in LB. For these specific LCRs to be detectable during animal colonization indicates that they have been selected under the *in vivo* conditions of the animals intestinal tract. For example, in colonisation experiments, the input strain 9000 confirmation (profile C/A_11b) was recovered in 8/12 isolates, with the remaining four having the large 1.4 Mb inversion as determined by PFGE (Fig. S6B). Intriguingly, the highest excretion level in that experiment was associated with an animal from which a strain with the inverted confirmation was recovered. Currently, there is no simple way to quantify the proportions of the

confirmations under specific conditions, with the exception of optical mapping for the isolates cultured in the laboratory.

As further support for these processes in cattle, extensive surveys of *E. coli* O157 in cattle herds [38, 39] determined that while the majority of isolates in any specific herd exhibit the same PFGE pattern, there are isolates with different profiles yet the same phage type (PT) [40, 41]. A recent study of persistent Lineage I strains isolated on a single farm also demonstrated that a 47.7 kbp deletion was a significant genomic difference between two of the strains [42]. We show that LCRs are the likely cause of the observed PFGE profile type expansion amongst PT21/28 bovine isolates in the UK. There has been one previous report of multiple deletions occurring during *E. coli* O157 colonisation of cattle, generating multiple PFGE types [43].

LCRs have been observed in a number of bacterial genera, including *Campylobacter*, *Yersinia*, *Staphylococcus* and *Salmonella* [22, 44-47]. For inversions the gene content and copy number is maintained but the prophage boundaries do change in composition and this could have an impact on prophage gene expression or the regulatory networks that they are part of [1]. A clear example of this was shown for *Campylobacter* where in one orientation the inversion completes an active prophage and in turn that provides resistance to certain infecting phages [44]. For *E. coli* O157 strain 9000 structural variants we measured a number of expression and phenotype changes, including motility and growth rate for variants with inversions. The most obvious differentials were present in the variant with a 220 kbp duplication. This included an increase in Stx expression, production and toxicity and a reduction in type 3 secretion. Our previous research has shown that Stx2a prophage integration into different *E. coli* backgrounds led to a repression of T3S, potentially via the CII protein [28]. Such cross-regulation would offer one pathway resulting in the concomitant reduction in T3S in the strain with the duplication.

Conclusions

We describe the first systematic genome structure comparison of strains across the main lineages *Escherichia coli* O157. LCRs, predominantly large inversions, were a common source genomic variation and appear to be generated by recombination between homologous prophage sequences. Importantly, we show that LCRs are generated during animal colonisation and laboratory culture and demonstrate that specific LCRs are associated with phenotypic changes. By definition, phenotypic heterogeneity is the occurrence of individuals within a genetically identical population that stochastically develop phenotypes of varying fitness within a homogenous environment [30, 32]. With the work presented here and that in other genera, it is evident that genome structural variants are a way to generate phenotypic heterogeneity in a clonal bacterial population and that relevant sub-populations can then be selected as conditions change in particular environments making it an important population survival strategy.

Materials and Methods

Bacterial strains and culture conditions

Bacterial strains and plasmids used in this study are listed in Table. S1. Bacteria were cultured in Luria-Bertani (LB) broth or M9 minimal media (Sigma- Aldrich) supplemented with 0.2% glucose, 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$. For TTSS expression bacteria were cultured overnight in LB and then inoculated into minimal essential medium (MEM)-HEPES (Sigma-Aldrich) supplemented with 0.1% glucose and 250 nM $Fe(NO_3)_3$. Antibiotics were used at the following concentrations when required: Chloramphenicol (50 µg/ml), Mitomycin C (2 µg/ml), Nalidixic acid (50 µg/ml).

PacBio Long-read sequencing

A total of 72 whole genome sequences, generated by PacBio long-read sequencing, were used for analysis in this study. The sequences of 31 strains were determined for this study and the remaining 41 were publicly available in the National Centre for Biotechnology Information (NCBI) database (Table S1).

Sequencing of the 31 isolates was conducted using a PacBio RS II long-read sequencing platform and carried out at the U. S. Department of Agriculture sequencing core facility in in Clay Center, Nebraska, USA. Qiagen Genomic-tip 100/G columns and a modified protocol, as previously described [48], were used to extract high molecular weight DNA. Using a g-TUBE (Corvaris), 10 µg of DNA was sheared to a targeted size of 20 kb and concentrated using 0.45x volume of AMPure PB magnetic beads (Pacific Biosciences). Following the manufacturer's protocol, 5 µg sheared DNA and the PacBio DNA SMRTbell Template Prep kit 1.0 were used to create the sequencing libraries. A BluePippin instrument (Sage Science) with the SMRTbell 15–20 kb setting was used to size select 10 kb or larger fragments. The library was bound with polymerase P5 and sequencing was conducted with the C3 chemistry and the 120 min data collection protocol. Individual libraries were constructed from some of the strain DNA preparations described above using an Illumina Nextera XT DNA sample preparation kits with appropriate indices tags according to the manufacturer's instructions (Illumina Inc., San Diego, CA). The libraries were pooled together and run on an Illumina MiSeq DNA sequencer (Illumina Inc., San Diego, CA). The genome of each strain was sequenced to a targeted depth of 50X coverage.

Genome assembly and annotation

SMRT analysis was used to generate a FASTQ file from the PacBio reads, which were then error-corrected using PBcR with self-correction [49]. The Celera Assembler was used to assemble the longest 20× coverage of the corrected reads. The resulting contigs were improved using Quiver [50] and annotation was conducted using a local instance of Do-It-Yourself Annotator (DIYA) [51]. Geneious (Biomatters) was used to remove duplicated sequence from the 5′ and 3′ ends to generate the circularized chromosome. To correct PacBio sequencing errors (homopolymers and SNPs), Illumina reads were mapped to the Quiver polished chromosome using Pilon [52]. Then, both PacBio

and Illumina reads were mapped to the Pilon-generated chromosome using Geneious Mapper. Additional sequencing errors were identified and corrected by manual editing in Geneious, resulting in a finished closed circularized chromosome. OriFinder was used to determine the origin of replication [53] and the chromosome was reoriented using the origin as base number one. Prophage regions were identified as described previously [9] using PHASTER [54].

MinION sequencing and SV read detection

Strain 9000 was sequenced by Oxford nanopore technologies MinION sequencing. High molecular weight genomic DNA was extracted from strain 9000 grown in LB (OD600 = 0.7) by standard phenol:chloroform extraction [55]. Genomic DNA was purified using Qiagen G100 Genomic Tips (Qiagen) with minor alterations including no vigorous mixing steps and final elution in 100µl of nuclease free water and quantified using a Qubit and the HS (high sensitivity) dsDNA assay kit (Thermofisher Scientific), following the manufacturer's instructions. Library preparation was performed using the Ligation kit SQK-LSK109 (Oxford Nanopore Technologies). The prepared libraries were loaded onto a FLO-MIN106 R9.4.1D flow cell (Oxford Nanopore Technologies) and sequenced using the MinION (Oxford Nanopore Technologies) for 72 h. Data produced in a raw FAST5 format was basecalled and de-multiplexed using Guppy v3.2.4 using the FAST protocol (Oxford Nanopore Technologies) into FASTQ format.

To identify if the Nanopore sequenced strain 9000 contained reads supporting multiple isoforms of the chromosome. Minimap2 v2.17 [56] and Samtools v1.7 [57] was used to align the Nanopore reads (removing secondary aligning reads) to samples Z1615, Z1723 and Z1767 each representing a different chromosomal isoform. Using Samtools v1.7 [57] and Bedtools v2.29.2 [58] reads were identified at either end of the each of the 5' and 3' breakpoints identified in those conformations. The number of reads that crossed each end of the 5' and 3' breakpoints for both conformations was calculated again using Samtools v1.7 [57] and Bedtools v2.29.2 [58].

Whole genome comparisons

Pairwise whole genome alignments were conducted with Easyfig [59] as described previously [9]. Genome .gbk files were modified so that prophage were represented as coloured blocks. AvrII restriction sites were identified in selected genomes using UGENE [60] and their loci were added to the respective genome .gbk files. Pairwise whole genome alignments between reference genomes from each lineage (9000, Sakai, TW14359 and 180) and each genome were performed using blastn [61] with the following parameters (-evalue 1e-10 -best_hit_score_edge 0.05 -best_hit_overhang 0.25 -perc_identity 70 -max_target_seqs 1 -outfmt 6). From the resulting alignment files, LCRs were identified within each genome by filtering all inverted homologous regions ≥ 50,000 bp relative to each reference strain.

Mapping homologous regions

Homologous regions within each genome were identified using blastn [61]. Blastn was performed on each individual genome using the same genome sequence as both reference and query with the following parameters (-evalue 1e-10 -best_hit_score_edge 0.05 -best_hit_overhang 0.25 -perc_identity 98 -max_target_seqs 1 -outfmt 6). Homologous regions that satisfied three conditions simultaneously were extracted from the blast output: (1) Homologous regions were ≥ 5000 bp (2) homologous regions ≥ 5000 bp were present in the genome at a frequency ≥ 2 (3) homologous regions were located before and after *dif* (terminus of replication). Equivalent analysis was repeated to determine homologous regions ≥ 8000 bp. Significant differences in the total number of repeats detected between lineages and the bias of repeat regions toward Ter was determined by one-way ANOVA with Dunnetts multiple comparisons test.

Circos plots [62] were used to visualise linked regions of homologous sequence within the genomes of selected strains. Custom circos input files were generated in which the data matrix was modified such that each circular genome was divided at prophage boundaries. Linked homologous regions and their sizes were determined using BLAST scores derived when querying a selected genome sequence to itself. Only BLAST hits with ≥ 98 % sequence homology and that were ≥ 5000 bp in length were included in the data matrix of circos input files. Within Circos plots the width of linked segments is proportional to the length of BLAST hits. Circos does not exactly map homology hits to linked chromosomal/prophage regions, instead connecting segments originate and end at the earliest available location within the linked region.

Phylogeny of 72 strains and PT21/28 strains

A core gene alignment was extracted from the fully assembled and annotated PacBio genomes of all 72 strains using ROARY [63] with parameters (-e -n -r -s -ap). The extracted multiple alignment was used to Maximum-liklihood phylogenetic trees FastTree [64] (-gtr) and trees were visualised with iTOL [65]. To determine the phylogenetic relationship of PT21/28 strains high quality illumine sequencing reads were mapped to the reference STEC O157 strain, Sakai (GenBank accession BA000007), using Burrows-Wheeler Aligner – Maximum Exact Matching (BWA MEM (v0.7.2)) [66]. The sequence alignment map output from BWA were sorted and indexed to produce a binary alignment map (BAM) using Samtools (v1.1) [67]. Genome Analysis Toolkit (GATK v2.6.5) was then used to create a variant call format (VCF) file from each of the sorted BAMs, which were further parsed to extract only SNP positions of high quality (mapping quality (MQ) > 30, depth (DP) > 10, variant ratio> 0.9). Hierarchical single linkage clustering was performed on the pairwise SNP difference between all isolates at descending distance thresholds (Δ250, Δ100, Δ50, Δ25, Δ10, Δ5, Δ0) [68]. SNP alignments were created tolerating positions where >80% of isolates had a base call with regions of recombination masked using Gubbins v2.0.0 [69]. Maximum likelihood phylogenies

were computed using IQ-TREE v2.0.4 [70] with the best-fit model automatically selected and near zero branches collapsed into polytomies.

Pulsed-field Gel Electrophoresis

All strains analysed by PFGE were cultured in LB or M9 medium at 37°C overnight with agitation. Genomic DNA was purified using the CHEF Bacterial Genomic DNA Plug Kit (Bio-Rad) according to manufacturer guidelines. DNA restriction digestion with AvrII (BlnI) (Takara) and subsequent PFGE was done according to the PulseNet O157 guidelines [71], using a CHEF-DR III system. *In silico* AvrII (BlnI) restriction digests of selected genomes was carried out in CLC Genomics Workbench (Qiagen).

RNA sequencing

Total RNA was extracted from three biological replicates of strains 9000, Z1615 Z1723, Z1767 using mirVana™ miRNA Isolation Kit (ThermoFisher) according to manufacturer guidelines.

Strains were cultured in either LB or M9 media to OD_{600} = 0.7. Ribosome depletion, cDNA library preparation and Illumina sequencing was carried out by Vertis Biotechnologie AG (Freising, Germany). Total RNA samples were purified and concentrated using the Agencourt RNAClean XP kit (Beckman Coulter Genomics) and the RNA integrity was assessed by capillary electrophoresis. Ribosomal RNA molecules were depleted using the Ribo-Zero rRNA Removal Kit for bacteria (Illumina). The ribodepleted RNA samples were first fragmented using ultrasound (4 pulses of 30 s each at 4°C) and oligonucleotide adapters were then ligated to the 3' end of the RNA molecules. First-strand cDNA synthesis was performed using M-MLV reverse transcriptase and the 3' adapter as primer. The first-strand cDNA was purified and the 5' Illumina TruSeq sequencing adapter was ligated to the 3' end of the antisense cDNA. The resulting cDNA was PCR-amplified to about 10-20 ng/µl using a high-fidelity DNA polymerase. The cDNA was purified using the Agencourt AMPure XP kit (Beckman Coulter Genomics) and was analyzed by capillary electrophoresis. Purified cDNA was pooled and sequenced on an Illumina NextSeq 500 system using 75 bp read length. RNA-sequencing reads were mapped to the strain 9000 reference genome (CP018252.1) using STAR 2.7.0e [72] with the following parameters (--quantMode GeneCounts and --sjdbGTFfeatureExon CDS). Prior to read mapping the reference strain 9000 was annotated using Prodigal version 2.6 [73]. The loci of previously identified *E. coli* O157 sRNA [7] were found in strain 9000 using BLASTn and manually added to strain 9000 .gtf file. Column 3 of the reference GTF file (feature) was manually modified to CDS for all genetic features. Differential expressed (DE) genes were identified with edgeR [74] (p-values =0.05) using the glmQLFit + glmQLFTest parameters. RNA-seq data was uploaded to NCBI Gene Expression Omnibus (GEO) (Accession: GSE158899).

Stx toxin ELISA

3 ml LB was inoculated directly from glycerol stocks and grown overnight at 37 °C. 6 ml LB was inoculated 1/100 from overnight cultures and grown to an OD_{600nm} = 0.6–0.8. Mitomycin C (2 µg/ml) was added and lysis allowed to proceed for 24 h. After 24 h, 1 ml culture was taken and live cells and cell debris removed by centrifugation (13,000 rpm). Stx toxin containing supernatants were further sterilized by syringe filtering (0.22 µm; Milipore). The level of Stx toxin in each sample was assayed using the RIDASCREEN® Verotoxin ELISA kit (R-Biopharm) according to manufacturer guidelines. Differences Stx2 production was assessed by ordinary one-way ANOVA with multiple comparisons where each strain was compared with Z1723.

Stx Vero cell toxicity

Cytotoxicity of Stx2 toxin was measured on Vero cell monolayers cultured in RPMI medium (Sigma-Aldrich). Cells (100 µl) were plated into 96-well microtitre plates and at ~ 75% confluence the culture medium was replaced with RPMI medium containing diluted (1:1000) Stx2 toxin supernatants. Vero cells were exposed to Stx2 toxin for 72 hours at 37°C, 5% CO_2. Surviving cells were fixed using paraformaldehyde (2 %) and stained with crystal violet (10 %). Crystal violet was solubilized with 10% acetic acid live/dead cells were quantified spectrophotometrically at 590 nm. Cells exposed to Triton X-100 (0.1 %) and RPMI were used as positive and negative controls for toxicity respectively. Strain toxicity was expressed as a percentage of the toxicity measured for RPMI control. Strain toxicity was analysed by ordinary one-way ANOVA with multiple comparisons where each strain was compared with Z1723.

Fitness assays

The fitness of strain 9000 variants grown in M9 media was calculated as described previously [75, 76]. Viable-cell counts for each competing strain were determined at time zero (t=0) and again after 24 h of co-culturing by selective plating. Fitness was calculated using the formula:

Fitness index (f.i.) = LN (N_i (1)/ N_i (0)) **/** LN (N_j (1)/ N_j (0)),

Where N_i (0) and N_i (1) = initial and final colony counts of strain Z1723 or 9000, respectively and

N_j (0) and N_j (1) = initial and final colony counts of structural variant strain (Z1766, Z1767 or Z1615), respectively

For controls WT strain 9000 or Z1723 were competed against Nal^r derivatives generated previously [25]. Fitness was analysed by ordinary one-way ANOVA with Dunnett's multiple comparisons test where each strain was compared with control.

RT-qPCR

Total RNA was extracted from cell pellets using a RNeasy® Mini kit (Qiagen) according to manufacturer guidelines. Extracted RNA was quantified and 2 µg of each samples was DNase treated using TURBO DNA-*free*™ kit. 200 ng of DNase treated RNA was then converted to cDNA using iScript™ Reverse Transcription Supermix (Bio-Rad) according to manufacturer guidelines. All qPCR reactions were carried out using iQ™ Syber® Green supermix (Bio-Rad) and *stx2a* specific primers (IDT-DNA): stx2a-F–GAAGAAGATGTTTATGGCGGTTT, stx2a-R–CCCGTCAACCTTCACTGTAA. Cycling conditions were: 95 °C for 15 s (1 cycle), 95 °C for 15 s; 60 °C for 1 min (40 cycles). Gene expression was quantified relative to a standard curve generated from Z1723 genomic DNA.

Optical mapping

Strains Z1723 and Z1767 were cultured from a single colony in LB medium to an OD_{600} = 0.7. 1 ml of cells/agarose plug were harvested (4000 g, 5 min) and intact chromosomes were extracted according to the Bionano Prep Cell Culture DNA Isolation Protocol (Bionano). Briefly, harvested cells were washed twice in Bionano Cell Buffer (Bionano). Washed cells were embedded in 2 % Low melt agarose plugs and cells were lysed (1 hr at 37°C) with lysozyme enzyme (100 µl) using CHEF Bacterial Genomic DNA Plug Kit (Bio-Rad). DNA containing plugs were washed twice with nuclease free water then treated with Proteinase K (Qiagen) in Bionano Lysis Buffer according to the Bionano Prep Cell Culture DNA Isolation Protocol. All subsequent procedure steps (RNase treatment, DNA extraction, quantitation and labelling) and optical mapping on Bionano Irys platform were provided as a service by Earlham Institute (Norwich, UK). Structural variant analysis was provided by Bionano and structural variant maps visualised using Bionano access (Bionano).

TTSS expression and secretion

Expression of *ler* and *sepL* was measured using GFP reporter fusion plasmids pDW-LEE1 [77] and pDW6 [78], respectively. Reporter plasmids were transformed into strains Z1723, Z1766 and Z1767 by electroporation and transformants were cultured overnight in LB media supplemented with chloramphenicol (50 µg/ml). Overnight cultures were diluted 1:100 into MEM-HEPES and grown at 37°C (200 rpm) to an OD_{600} 0.8 – 1.0. GFP fluorescence of 200 µL aliquots was measured in a 96-well blank microtiter plate using a FLUOstar Optima plate reader (BMG, Germany). The Gfp promoter-less plasmid pKC26 was used as a control [79].

For EspD secretion, bacteria were cultured in 50 ml of MEM-HEPES at 37°C (200 rpm) to an OD_{600} of 0.8–1.0. Bacterial cells were pelleted by centrifugation at 4000 *g* for 20 min, and supernatants were passed through low protein binding filters (0.45 µm). 10% TCA was used to precipitate proteins overnight, which were separated by centrifugation at 4000 *g* for 30 min at 4°C. The proteins were

suspended in 150 µl of 1.5 M Tris (pH 8.8). For bacterial lysates, bacterial pellets were suspended directly in SDS PAGE loading buffer. Proteins were separated by SDS-PAGE using standard methods and Western blotting performed as described previously for EspD and RecA [28].

Acknowledgements: The authors would like to thank Sandy Fryda-Bradley and the USMARC core sequencing facility for excellent technical assistance. The mention of a trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by the USDA and does not imply approval to the exclusion of other products that might be suitable. The USDA is an equal opportunity employer and provider.

Figures and Tables

Figure 1. Position of major chromosomal rearrangements in *E. coli* O157 genomes.
The relative positions of all LCRs ≥ 50 kb (blue lines) are marked on the chromosomal maps (grey line) of strains from Lineage 1c (purple), Lineage 1a (blue) Lineage I/II (orange) and Lineage II (green). Chromosomes are centred by the replication terminus (Ter), beginning and ending at the origin of replication (OriC). LCRs are shown relative to four reference strains: (**A**) 9000, Lineage 1c; (**B**) Sakai, Lineage Ia; (**C**) TW14359, Lineage I/II; (**D**) 180, Lineage II. The position of the main prophage (red line) are mapped for each comparison strain.

Figure 2. Mapping homologous regions (≥ 5000 bp) in *E. coli* O157.
The loci of all regions of homology ≥ 5000 bp (black/red) that are present as ≥ 2 copies per genome are mapped on the chromosomes (grey line) of strains from Lineage 1c (purple), Lineage 1a (blue) Lineage I/II (orange) and Lineage II (green) (**A**). The directions 5' – 3' of homologous sequences relative to OriC are shown with black indicating the inverse direction to red. Circos plots for Lineage Ic strain 9000 (**B**) and Lineage I/II strain 272 (**C**) show paired regions of homology. Prophage loci (red blocks) are shown on the respective circular genome maps (blue). Paired homologous regions are joined by arches: Chromosomal (blue) and within prophage (red).

Figure 3. Distinct PFGE restriction patterns of *E coli* O157 PT21/28 strains are largely accounted for by LCRs. (**A**) Phylogenetic distribution of Lineage Ic PT21/28 strains. The source attribution, human (red) or bovine (blue) for each strain and PFGE variation across the lineage (coloured blocks) are shown. (**B**) Pairwise whole genome comparison of ten PT21/28 strains with different PFGE profiles. Whole genomes (black lines) are centred by the replication terminus (Ter) and loci of prophage (yellow boxes), Stx prophage (ΦStx2c;blue and ΦStx2a;red) and AvrII sites (blue triangles) are shown. Direct (purple) and inverted (orange) homology at a blast cut-off of 10,000 bp between strains are plotted. (**C**) PFGE profile of the ten selected PT21/28 strains following AvrII

digestion. (**D**) *In silico* generated AvrII digestion pattern of the PacBio-generated sequences for each strain.

Figure 4. Detection of LCRs in *E. coli* O157 PT21/28 strain 9000 variants analysed from cattle colonization studies. Pairwise whole genome comparisons of strains from Trial 1 (**A**) and Trial 2 (**B**) are shown with direct (purple) and inverted (orange) homology at a blast cut-off of 10,000 bp between strains. Whole genomes (black lines) are centred by the replication terminus (Ter) and the loci of prophage (yellow boxes) and Stx prophage (ΦStx2c;blue and ΦStx2a;red) in each strain are mapped. (**C**) Circos plots showing the identified 220 kbp duplication in Z1723 (left) and 1.2 Mbp inversion in Z1767 (right) relative to progenitor strain 9000. Outer ring: Strain 9000 (grey) and LCR derivatives Z1723 and Z1767 (black); Middle ring: Loci of prophage (black) and prophage a LCR boundaries (red); Inner ring: GC content.

Figure 5. Optical mapping of *E. coli* O157 PT21/28 strain 9000 variant Z1723. Structural variant (SV) analysis identified a 220 Kb duplication (**A**) and 1.2 Mb inversion (**B**) in the population of Z1723 relative to the reference strain 9000. The genome map (orange) of reference strain 9000 and each Z1723 structural variant (green) are shown. Paired restriction sites (blue lines) are aligned between the reference and variant maps (grey lines). Unpaired restriction sites (purple lines) outside aligned regions are also shown. The SV map containing the 220 Kb duplication has been aligned to two reference strain 9000 genome maps to demonstrate the hybrid composition of the map containing ΦStx2c at 3.4 Mb and an inverted 220 Kb duplicated region originating from between 2.2 and 2.4 Mb.

Figure 6. Shiga toxin expression, production and toxicity of Strain 9000 structural variants. The chromosomal location of all differentially expressed genes in Z1723 (orange bars) are mapped to the reference strain 9000 genome (**A**). Prophage (red blocks), the 220 kb duplication and ΦStx2a regions are highlighted. Expression of stx2a (**B**) total Stx toxin production (**C**) and Vero cell toxicity Stx (**D**) was measured for Trial 2 strains Z1723, Z1766 and Z1767 in M9 media. Mean values +/- SEM of four biological replicates (n = 4) are shown for each assay. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; ****$p \leq 0.0001$

Figure 7. Type III secretion, competitive fitness and motility phenotypes of strain 9000 structural variants. **(A)** Expression of the LEE master regulator *ler* and LEE4 chaperone *sepL* was measured by Gfp reporter fusions (n = 3). **(B)** Detection of the LEE effector EspD in the culture supernatants of each strain by Western blot (n = 3). Corresponding cellular RecA levels were used as a control.

(C) Competitive fitness of strains after 24 h co-culturing in M9 media (n = 6). **(D)** motility of strains after 6 h on Tryptone swarm plates (n = 20). Mean values +/- SEM are shown for each assay. * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; ****$p \leq 0.0001$

Supplementary data information

Eight Supplementary figures with legends

Three supplementary Tables:

References

1. Brussow H, Canchaya C, Hardt WD. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev. 2004;68(3):560-602, table of contents. Epub 2004/09/09. doi: 10.1128/MMBR.68.3.560-602.2004. PubMed PMID: 15353570; PubMed Central PMCID: PMCPMC515249.
2. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. The Diverse Impacts of Phage Morons on Bacterial Fitness and Virulence. Adv Virus Res. 2019;103:1-31. Epub 2019/01/13. doi: 10.1016/bs.aivir.2018.08.001. PubMed PMID: 30635074.
3. Argov T, Azulay G, Pasechnek A, Stadnyuk O, Ran-Sapir S, Borovok I, et al. Temperate bacteriophages as regulators of host behavior. Curr Opin Microbiol. 2017;38:81-7. Epub 2017/05/26. doi: 10.1016/j.mib.2017.05.002. PubMed PMID: 28544996.
4. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. ISME J. 2016;10(12):2854-66. Epub 2016/06/04. doi: 10.1038/ismej.2016.79. PubMed PMID: 27258950; PubMed Central PMCID: PMCPMC5148200.
5. Davies EV, Winstanley C, Fothergill JL, James CE. The role of temperate bacteriophages in bacterial infection. FEMS Microbiol Lett. 2016;363(5):fnw015. Epub 2016/01/31. doi: 10.1093/femsle/fnw015. PubMed PMID: 26825679.
6. Wang X, Kim Y, Ma Q, Hong SH, Pokusaeva K, Sturino JM, et al. Cryptic prophages help bacteria cope with adverse environments. Nat Commun. 2010;1:147. Epub 2011/01/27. doi: 10.1038/ncomms1146. PubMed PMID: 21266997; PubMed Central PMCID: PMCPMC3105296.
7. Tree JJ, Granneman S, McAteer SP, Tollervey D, Gally DL. Identification of bacteriophage-encoded anti-sRNAs in pathogenic Escherichia coli. Molecular cell. 2014;55(2):199-213. doi: 10.1016/j.molcel.2014.05.006. PubMed PMID: 24910100; PubMed Central PMCID: PMC4104026.
8. Ferens WA, Hovde CJ. Escherichia coli O157:H7: animal reservoir and sources of human infection. Foodborne pathogens and disease. 2011;8(4):465-87. Epub 2010/12/02. doi: 10.1089/fpd.2010.0673. PubMed PMID: 21117940; PubMed Central PMCID: PMCPMC3123879.
9. Shaaban S. C, LA., McAteer, SP., Jenkins, C., Dallman, TJ., Bono, JL., and Gally, DL. . Evolution of a zoonotic pathogen: investigating prophage diversity in enterohaemorrhagic Escherichia coli O157 by long-read sequencing. Microbial Genomics. 2016. doi: 10.1099/mgen.0.000096.
10. Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, et al. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature. 2001;409(6819):529-33. doi: 10.1038/35054089. PubMed PMID: 11206551.
11. Dallman T, Ashton., PM., Byrne, L., Perry, NT., Petrovska, L., Ellis, R., Allison, L., Hanson M., Holmes, A., Gunn, GJ., Chase-Topping, ME., Woolhouse MEJ, Grant KA, Gally, DL, Wain, J., and Jenkins, C.. Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. Microbial Genomics. 2015. doi: 10.1099/mgen.0.000029.
12. Heiman KE, Mody RK, Johnson SD, Griffin PM, Gould LH. Escherichia coli O157 Outbreaks in the United States, 2003-2012. Emerging infectious diseases. 2015;21(8):1293-301. doi: 10.3201/eid2108.141364. PubMed PMID: 26197993; PubMed Central PMCID: PMCPMC4517704.
13. Karmali MA. Host and pathogen determinants of verocytotoxin-producing Escherichia coli-associated hemolytic uremic syndrome. Kidney Int Suppl. 2009;(112):S4-7. doi: 10.1038/ki.2008.608. PubMed PMID: 19180132.
14. Obrig TG, Karpman D. Shiga toxin pathogenesis: kidney complications and renal failure. Current topics in microbiology and immunology. 2012;357:105-36. doi: 10.1007/82_2011_172. PubMed PMID: 21983749; PubMed Central PMCID: PMC3779650.
15. Brandal LT, Wester AL, Lange H, Lobersli I, Lindstedt BA, Vold L, et al. Shiga toxin-producing escherichia coli infections in Norway, 1992-2012: characterization of isolates and identification of risk factors for haemolytic uremic syndrome. BMC infectious diseases. 2015;15:324.

doi: 10.1186/s12879-015-1017-6. PubMed PMID: 26259588; PubMed Central PMCID: PMC4531490.

16. Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, et al. Phylogenetic Clades 6 and 8 of Enterohemorrhagic Escherichia coli O157:H7 With Particular stx Subtypes are More Frequently Found in Isolates From Hemolytic Uremic Syndrome Patients Than From Asymptomatic Carriers. Open Forum Infect Dis. 2014;1(2):ofu061. doi: 10.1093/ofid/ofu061. PubMed PMID: 25734131; PubMed Central PMCID: PMCPMC4281788.

17. Buvens G, De Gheldre Y, Dediste A, de Moreau AI, Mascart G, Simon A, et al. Incidence and virulence determinants of verocytotoxin-producing Escherichia coli infections in the Brussels-Capital Region, Belgium, in 2008-2010. Journal of clinical microbiology. 2012;50(4):1336-45. doi: 10.1128/JCM.05317-11. PubMed PMID: 22238434; PubMed Central PMCID: PMC3318570.

18. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. Shiga toxin subtypes display dramatic differences in potency. Infection and immunity. 2011;79(3):1329-37. doi: 10.1128/IAI.01182-10. PubMed PMID: 21199911; PubMed Central PMCID: PMC3067513.

19. Iguchi A, Iyoda S, Terajima J, Watanabe H, Osawa R. Spontaneous recombination between homologous prophage regions causes large-scale inversions within the Escherichia coli O157:H7 chromosome. Gene. 2006;372:199-207. doi: 10.1016/j.gene.2006.01.005. PubMed PMID: 16516407.

20. Kotewicz ML, Jackson SA, LeClerc JE, Cebula TA. Optical maps distinguish individual strains of Escherichia coli O157 : H7. Microbiology. 2007;153(Pt 6):1720-33. doi: 10.1099/mic.0.2006/004507-0. PubMed PMID: 17526830.

21. Periwal V, Scaria V. Insights into structural variations and genome rearrangements in prokaryotic genomes. Bioinformatics. 2015;31(1):1-9. Epub 2014/09/06. doi: 10.1093/bioinformatics/btu600. PubMed PMID: 25189783.

22. Hughes D. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. Genome Biol. 2000;1(6):REVIEWS0006. doi: 10.1186/gb-2000-1-6-reviews0006. PubMed PMID: 11380986; PubMed Central PMCID: PMCPMC138892.

23. Darmon E, Leach DR. Bacterial genome instability. Microbiol Mol Biol Rev. 2014;78(1):1-39. Epub 2014/03/07. doi: 10.1128/MMBR.00035-13. PubMed PMID: 24600039; PubMed Central PMCID: PMCPMC3957733.

24. Ooka T, Ogura Y, Asadulghani M, Ohnishi M, Nakayama K, Terajima J, et al. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. Genome research. 2009;19(10):1809-16. Epub 2009/07/01. doi: 10.1101/gr.089615.108. PubMed PMID: 19564451; PubMed Central PMCID: PMCPMC2765283.

25. Fitzgerald SF, Beckett, AE., Palarea-Albaladejo, J., McAteer, S., Shaaban, S., Morgan, J., Ahmad, NI., Young, R., Mabbott, NA., Bono, JL., Gally, DL., McNeilly, TN. Shiga toxin sub-type 2a increases the efficiency of Escherichia coli O157 transmission between animals and restricts epithelial regeneration in bovine enteroids PLoS pathogens. 2019. Epub 03/10/19. doi: 10.1371/journal.ppat.1008003

26. Corbishley A, Ahmad NI, Hughes K, Hutchings MR, McAteer SP, Connelley TK, et al. Strain-dependent cellular immune responses in cattle following Escherichia coli O157:H7 colonization. Infection and immunity. 2014;82(12):5117-31. doi: 10.1128/IAI.02462-14. PubMed PMID: 25267838; PubMed Central PMCID: PMC4249286.

27. Taylor DE, Rooker M, Keelan M, Ng LK, Martin I, Perna NT, et al. Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic Escherichia coli O157:H7 isolates. J Bacteriol. 2002;184(17):4690-8. Epub 2002/08/10. doi: 10.1128/jb.184.17.4690-4698.2002. PubMed PMID: 12169592; PubMed Central PMCID: PMCPMC135296.

28. Xu X, McAteer SP, Tree JJ, Shaw DJ, Wolfson EB, Beatson SA, et al. Lysogeny with Shiga toxin 2-encoding bacteriophages represses type III secretion in enterohemorrhagic Escherichia coli. PLoS pathogens. 2012;8(5):e1002672. doi: 10.1371/journal.ppat.1002672. PubMed PMID: 22615557; PubMed Central PMCID: PMCPMC3355084.

29. Cowley LA, Dallman TJ, Fitzgerald S, Irvine N, Rooney PJ, McAteer SP, et al. Short-term evolution of Shiga toxin-producing Escherichia coli O157:H7 between two food-borne outbreaks. Microb Genom. 2016;2(9):e000084. doi: 10.1099/mgen.0.000084. PubMed PMID: 28348875; PubMed Central PMCID: PMCPMC5320650.
30. Ackermann M. A functional perspective on phenotypic heterogeneity in microorganisms. Nature reviews Microbiology. 2015;13(8):497-508. doi: 10.1038/nrmicro3491. PubMed PMID: 26145732.
31. Martins BM, Locke JC. Microbial individuality: how single-cell heterogeneity enables population level strategies. Curr Opin Microbiol. 2015;24:104-12. doi: 10.1016/j.mib.2015.01.003. PubMed PMID: 25662921.
32. Grimbergen AJ, Siebring J, Solopova A, Kuipers OP. Microbial bet-hedging: the power of being different. Curr Opin Microbiol. 2015;25:67-72. doi: 10.1016/j.mib.2015.04.008. PubMed PMID: 26025019.
33. Zhang Z, Du C, de Barsy F, Liem M, Liakopoulos A, van Wezel GP, et al. Antibiotic production in Streptomyces is organized by a division of labor through terminal genomic differentiation. Sci Adv. 2020;6(3):eaay5781. doi: 10.1126/sciadv.aay5781. PubMed PMID: 31998842; PubMed Central PMCID: PMCPMC6962034.
34. Lesterlin C, Pages C, Dubarry N, Dasgupta S, Cornet F. Asymmetry of chromosome Replichores renders the DNA translocase activity of FtsK essential for cell division and cell shape maintenance in Escherichia coli. PLoS Genet. 2008;4(12):e1000288. Epub 2008/12/06. doi: 10.1371/journal.pgen.1000288. PubMed PMID: 19057667; PubMed Central PMCID: PMCPMC2585057.
35. Bobay LM, Touchon M, Rocha EP. Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. PLoS Genet. 2013;9(9):e1003825. doi: 10.1371/journal.pgen.1003825. PubMed PMID: 24086157; PubMed Central PMCID: PMCPMC3784561.
36. De Paepe M, Hutinet G, Son O, Amarir-Bouhram J, Schbath S, Petit MA. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. PLoS Genet. 2014;10(3):e1004181. Epub 2014/03/08. doi: 10.1371/journal.pgen.1004181. PubMed PMID: 24603854; PubMed Central PMCID: PMCPMC3945230.
37. Asadulghani M, Ogura Y, Ooka T, Itoh T, Sawaguchi A, Iguchi A, et al. The defective prophage pool of Escherichia coli O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. PLoS pathogens. 2009;5(5):e1000408. Epub 2009/05/05. doi: 10.1371/journal.ppat.1000408. PubMed PMID: 19412337; PubMed Central PMCID: PMCPMC2669165.
38. Henry MK, Tongue SC, Evans J, Webster C, Mc KI, Morgan M, et al. British Escherichia coli O157 in Cattle Study (BECS): to determine the prevalence of E. coli O157 in herds with cattle destined for the food chain. Epidemiology and infection. 2017;145(15):3168-79. doi: 10.1017/S0950268817002151. PubMed PMID: 28925340.
39. Matthews L, McKendrick IJ, Ternent H, Gunn GJ, Synge B, Woolhouse ME. Super-shedding cattle and the transmission dynamics of Escherichia coli O157. Epidemiology and infection. 2006;134(1):131-42. doi: 10.1017/S0950268805004590. PubMed PMID: 16409660; PubMed Central PMCID: PMC2870353.
40. Pearce MC, Jenkins C, Vali L, Smith AW, Knight HI, Cheasty T, et al. Temporal shedding patterns and virulence factors of Escherichia coli serogroups O26, O103, O111, O145, and O157 in a cohort of beef calves and their dams. Appl Environ Microbiol. 2004;70(3):1708-16. doi: 10.1128/aem.70.3.1708-1716.2004. PubMed PMID: 15006796; PubMed Central PMCID: PMCPMC368277.
41. Vali L, Wisely KA, Pearce MC, Turner EJ, Knight HI, Smith AW, et al. High-level genotypic variation and antibiotic sensitivity among Escherichia coli O157 strains isolated from two Scottish

beef cattle farms. Appl Environ Microbiol. 2004;70(10):5947-54. doi: 10.1128/AEM.70.10.5947-5954.2004. PubMed PMID: 15466537; PubMed Central PMCID: PMCPMC522067.

42. Stanton E, Wahlig TA, Park D, Kaspar CW. Chronological set of E. coli O157:H7 bovine strains establishes a role for repeat sequences and mobile genetic elements in genome diversification. BMC genomics. 2020;21(1):562. Epub 2020/08/19. doi: 10.1186/s12864-020-06943-x. PubMed PMID: 32807088; PubMed Central PMCID: PMCPMC7430833.

43. Yoshii N, Ogura Y, Hayashi T, Ajiro T, Sameshima T, Nakazawa M, et al. pulsed-field gel electrophoresis profile changes resulting from spontaneous chromosomal deletions in enterohemorrhagic Escherichia coli O157:H7 during passage in cattle. Appl Environ Microbiol. 2009;75(17):5719-26. doi: 10.1128/AEM.00558-09. PubMed PMID: 19581472; PubMed Central PMCID: PMCPMC2737899.

44. Scott AE, Timms AR, Connerton PL, Loc Carrillo C, Adzfa Radzum K, Connerton IF. Genome dynamics of Campylobacter jejuni in response to bacteriophage predation. PLoS pathogens. 2007;3(8):e119. doi: 10.1371/journal.ppat.0030119. PubMed PMID: 17722979; PubMed Central PMCID: PMCPMC1950947.

45. Darling AE, Miklos I, Ragan MA. Dynamics of genome rearrangement in bacterial populations. PLoS Genet. 2008;4(7):e1000128. doi: 10.1371/journal.pgen.1000128. PubMed PMID: 18650965; PubMed Central PMCID: PMCPMC2483231.

46. Guerillot R, Kostoulias X, Donovan L, Li L, Carter GP, Hachani A, et al. Unstable chromosome rearrangements in Staphylococcus aureus cause phenotype switching associated with persistent infections. Proc Natl Acad Sci U S A. 2019;116(40):20135-40. doi: 10.1073/pnas.1904861116. PubMed PMID: 31527262; PubMed Central PMCID: PMCPMC6778178.

47. Sun S, Ke R, Hughes D, Nilsson M, Andersson DI. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. PloS one. 2012;7(8):e42639. doi: 10.1371/journal.pone.0042639. PubMed PMID: 22880062; PubMed Central PMCID: PMCPMC3411829.

48. Clawson ML, Keen JE, Smith TP, Durso LM, McDaneld TG, Mandrell RE, et al. Phylogenetic classification of Escherichia coli O157:H7 strains of human and bovine origin using a novel set of nucleotide polymorphisms. Genome Biol. 2009;10(5):R56. doi: 10.1186/gb-2009-10-5-r56. PubMed PMID: 19463166; PubMed Central PMCID: PMCPMC2718522.

49. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, McVey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013;14(9):R101. doi: 10.1186/gb-2013-14-9-r101. PubMed PMID: 24034426; PubMed Central PMCID: PMCPMC4053942.

50. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10(6):563-9. doi: 10.1038/nmeth.2474. PubMed PMID: 23644548.

51. Stewart AC, Osborne B, Read TD. DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics. 2009;25(7):962-3. doi: 10.1093/bioinformatics/btp097. PubMed PMID: 19254921; PubMed Central PMCID: PMCPMC2660880.

52. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PloS one. 2014;9(11):e112963. Epub 2014/11/20. doi: 10.1371/journal.pone.0112963. PubMed PMID: 25409509; PubMed Central PMCID: PMCPMC4237348.

53. Luo H, Zhang CT, Gao F. Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. Front Microbiol. 2014;5:482. doi: 10.3389/fmicb.2014.00482. PubMed PMID: 25309521; PubMed Central PMCID: PMCPMC4164010.

54. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 2016;44(W1):W16-21. Epub 2016/05/05. doi: 10.1093/nar/gkw387. PubMed PMID: 27141966; PubMed Central PMCID: PMCPMC4987931.

55. Green MR, Sambrook J. Isolation and quantification of DNA. Cold Spring Harbor Protocols. 2018;2018(6):pdb. top093336.
56. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100. Epub 2018/05/12. doi: 10.1093/bioinformatics/bty191. PubMed PMID: 29750242; PubMed Central PMCID: PMCPMC6137996.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. Epub 2009/06/10. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PubMed Central PMCID: PMCPMC2723002.
58. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2. Epub 2010/01/30. doi: 10.1093/bioinformatics/btq033. PubMed PMID: 20110278; PubMed Central PMCID: PMCPMC2832824.
59. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009-10. Epub 2011/02/01. doi: 10.1093/bioinformatics/btr039. PubMed PMID: 21278367; PubMed Central PMCID: PMCPMC3065679.
60. Okonechnikov K, Golosova O, Fursov M, team U. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166-7. Epub 2012/03/01. doi: 10.1093/bioinformatics/bts091. PubMed PMID: 22368248.
61. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PubMed PMID: 2231712.
62. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009;19(9):1639-45. Epub 2009/06/23. doi: 10.1101/gr.092759.109. PubMed PMID: 19541911; PubMed Central PMCID: PMCPMC2752132.
63. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PubMed Central PMCID: PMC4817141.
64. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one. 2010;5(3):e9490. doi: 10.1371/journal.pone.0009490. PubMed PMID: 20224823; PubMed Central PMCID: PMCPMC2835736.
65. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47(W1):W256-W9. Epub 2019/04/02. doi: 10.1093/nar/gkz239. PubMed PMID: 30931475; PubMed Central PMCID: PMCPMC6602468.
66. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. doi: 10.1093/bioinformatics/btp698. PubMed PMID: 20080505; PubMed Central PMCID: PMCPMC2828108.
67. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMCPMC2928508.
68. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. Bioinformatics. 2018;34(17):3028-9. doi: 10.1093/bioinformatics/bty212. PubMed PMID: 29659710.
69. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15. doi: 10.1093/nar/gku1196. PubMed PMID: 25414349; PubMed Central PMCID: PMCPMC4330336.
70. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530-4. doi: 10.1093/molbev/msaa015. PubMed PMID: 32011700; PubMed Central PMCID: PMCPMC7182206.

71. Ribot EM, Fair MA, Gautom R, Cameron DN, Hunter SB, Swaminathan B, et al. Standardization of pulsed-field gel electrophoresis protocols for the subtyping of Escherichia coli O157:H7, Salmonella, and Shigella for PulseNet. Foodborne pathogens and disease. 2006;3(1):59-67. Epub 2006/04/11. doi: 10.1089/fpd.2006.3.59. PubMed PMID: 16602980.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMCPMC3530905.
73. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. doi: 10.1186/1471-2105-11-119. PubMed PMID: 20211023; PubMed Central PMCID: PMCPMC2848648.
74. Nikolayeva O, Robinson MD. edgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. Methods Mol Biol. 2014;1150:45-79. doi: 10.1007/978-1-4939-0512-6_3. PubMed PMID: 24743990.
75. Dykhuizen DE. Experimental Studies of Natural Selection in Bacteria. Annual Review of Ecology and Systematics. 1990;21(1):373-98. doi: 10.1146/annurev.es.21.110190.002105.
76. Lenski RE. Quantifying fitness and gene stability in microorganisms. Biotechnology. 1991;15:173-92. Epub 1991/01/01. doi: 10.1016/b978-0-409-90199-3.50015-2. PubMed PMID: 2009380.
77. Fernandez-Brando RJ, Yamaguchi N, Tahoun A, McAteer SP, Gillespie T, Wang D, et al. Type III Secretion-Dependent Sensitivity of Escherichia coli O157 to Specific Ketolides. Antimicrob Agents Chemother. 2016;60(1):459-70. Epub 2015/11/04. doi: 10.1128/AAC.02085-15. PubMed PMID: 26525795; PubMed Central PMCID: PMCPMC4704242.
78. Wang D, Roe AJ, McAteer S, Shipston MJ, Gally DL. Hierarchal type III secretion of translocators and effectors from Escherichia coli O157:H7 requires the carboxy terminus of SepL that binds to Tir. Mol Microbiol. 2008;69(6):1499-512. Epub 2008/08/05. doi: 10.1111/j.1365-2958.2008.06377.x. PubMed PMID: 18673458.
79. Holden N, Totsika M, Dixon L, Catherwood K, Gally DL. Regulation of P-fimbrial phase variation frequencies in Escherichia coli CFT073. Infection and immunity. 2007;75(7):3325-34. Epub 2007/04/25. doi: 10.1128/IAI.01989-06. PubMed PMID: 17452474; PubMed Central PMCID: PMCPMC1932927.

Figure 1.

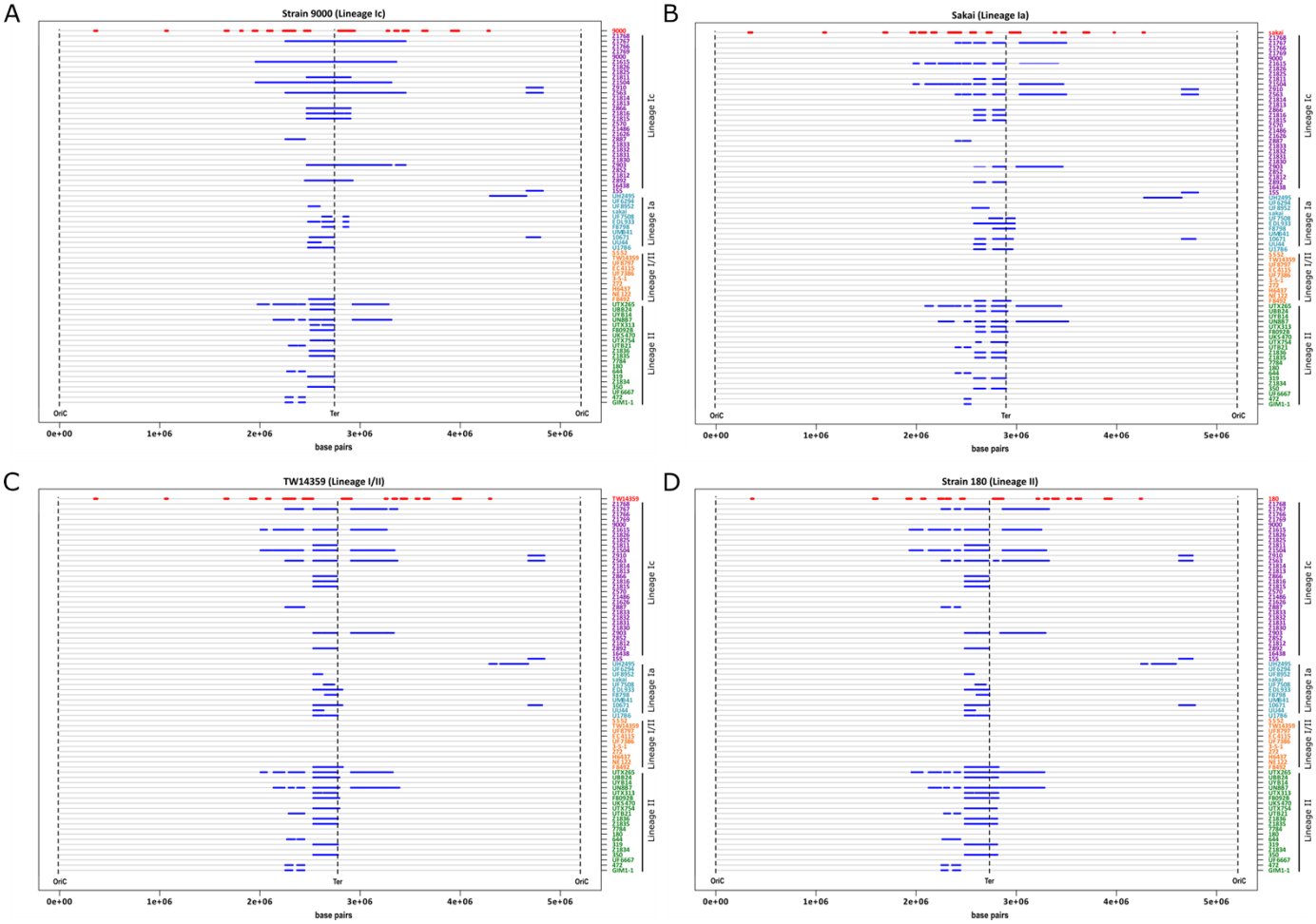

Figure 2.

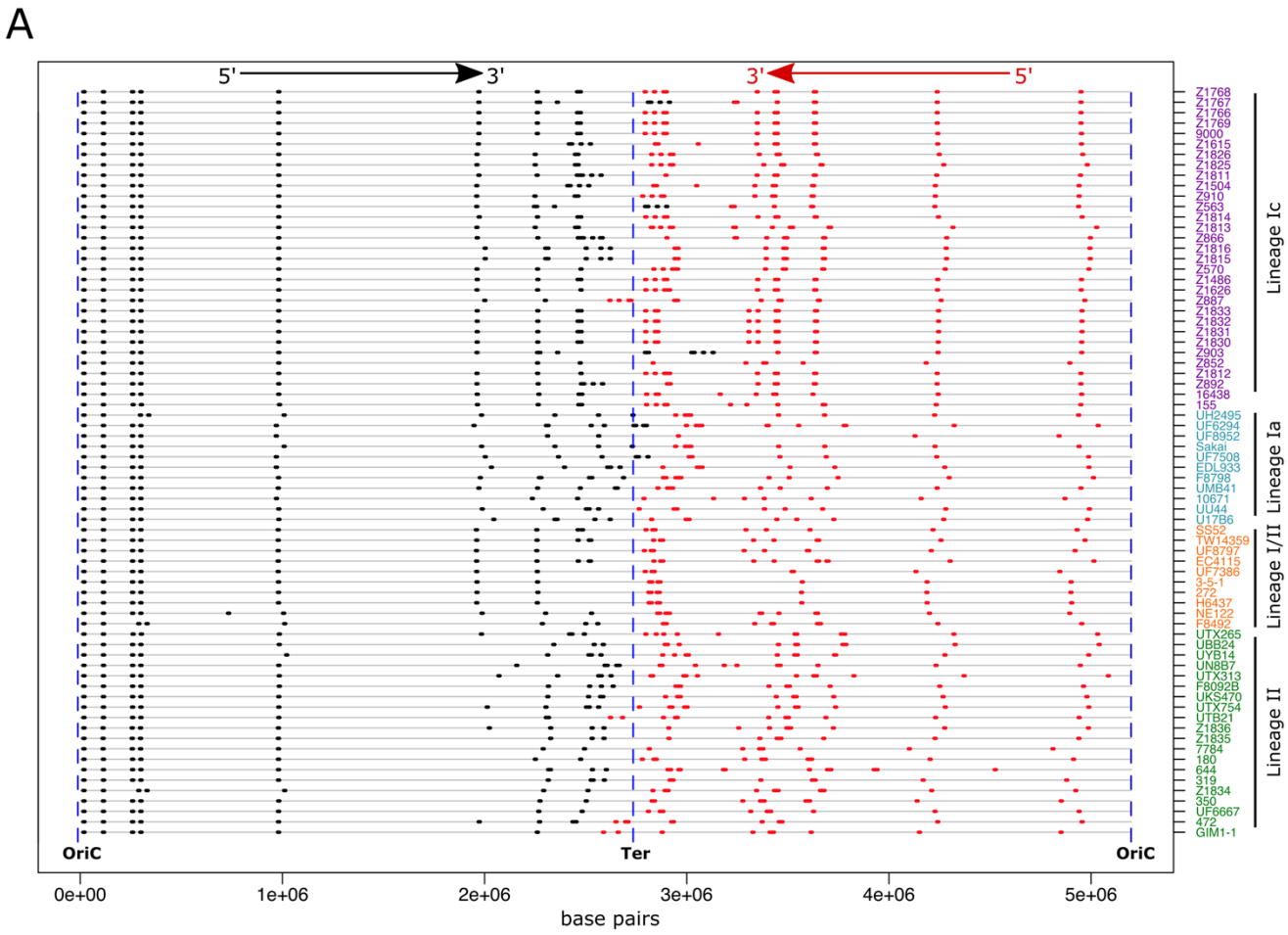

Figure 3.

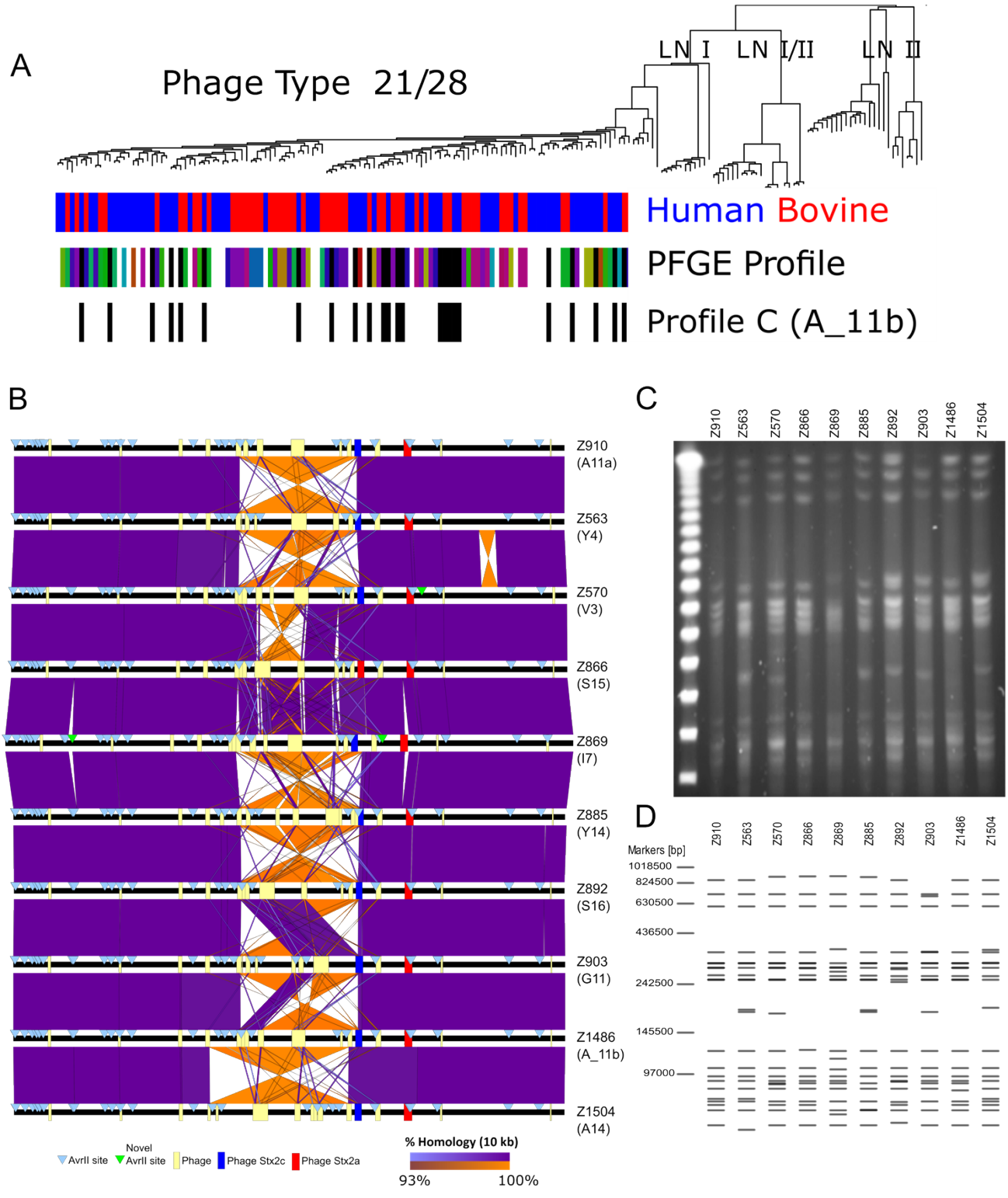

Figure 4.

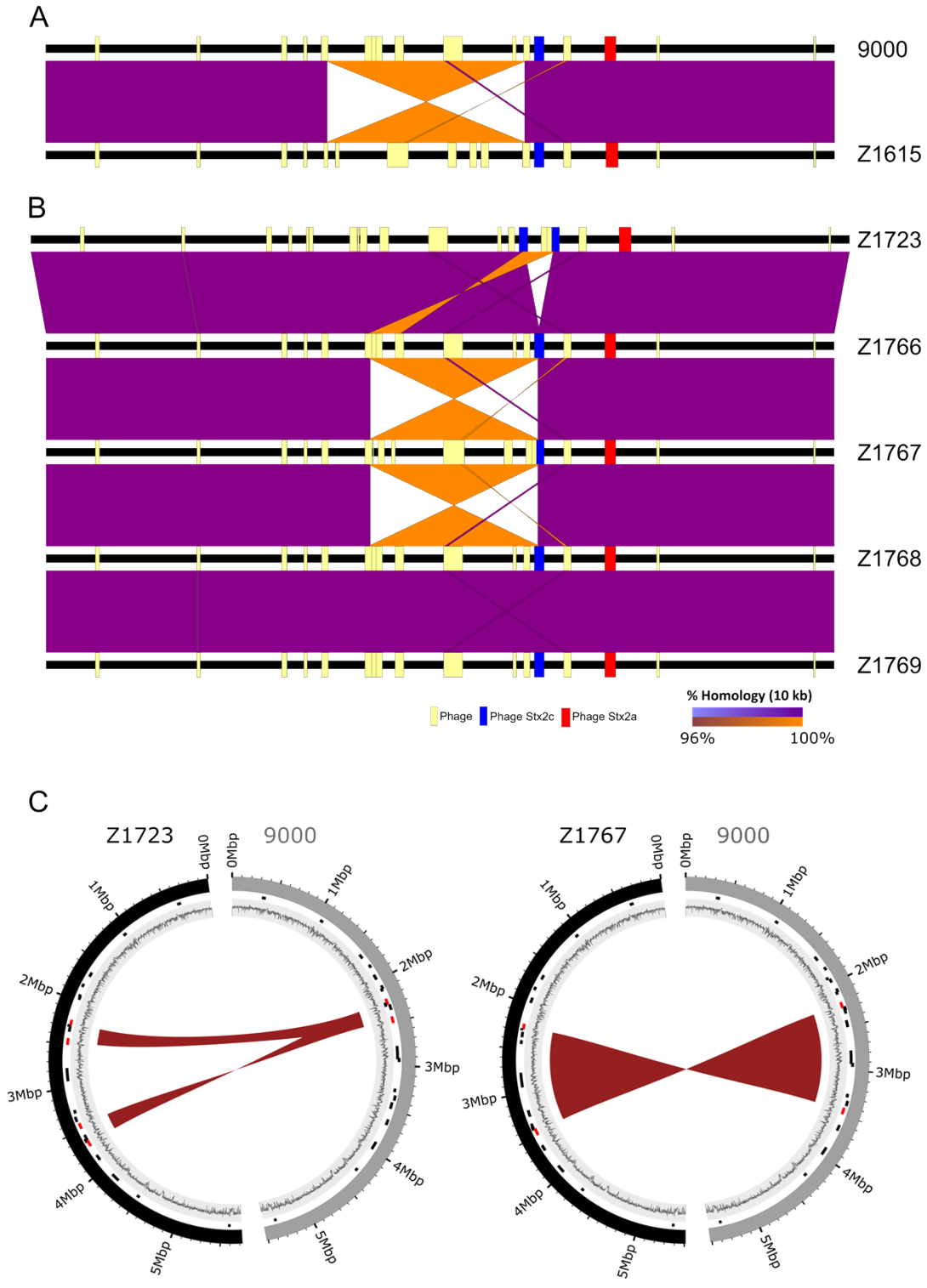

Figure 5.

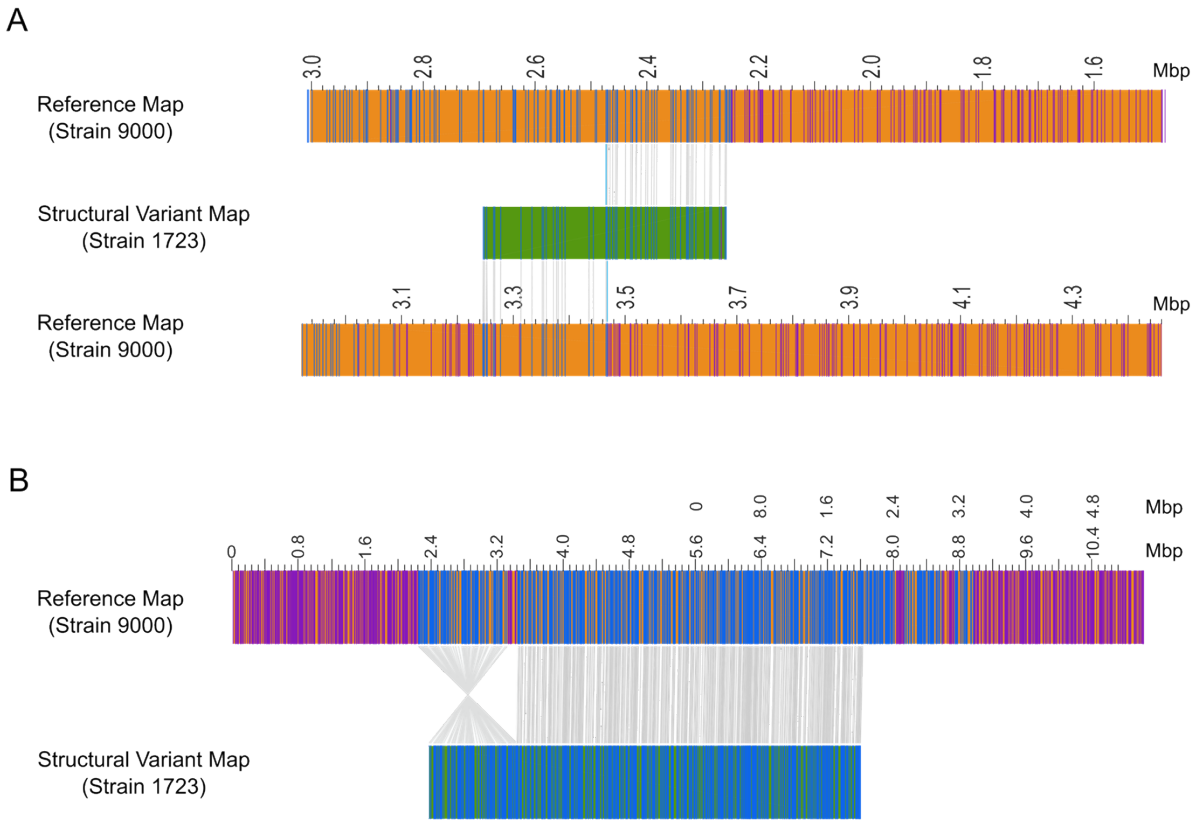

Figure 6.

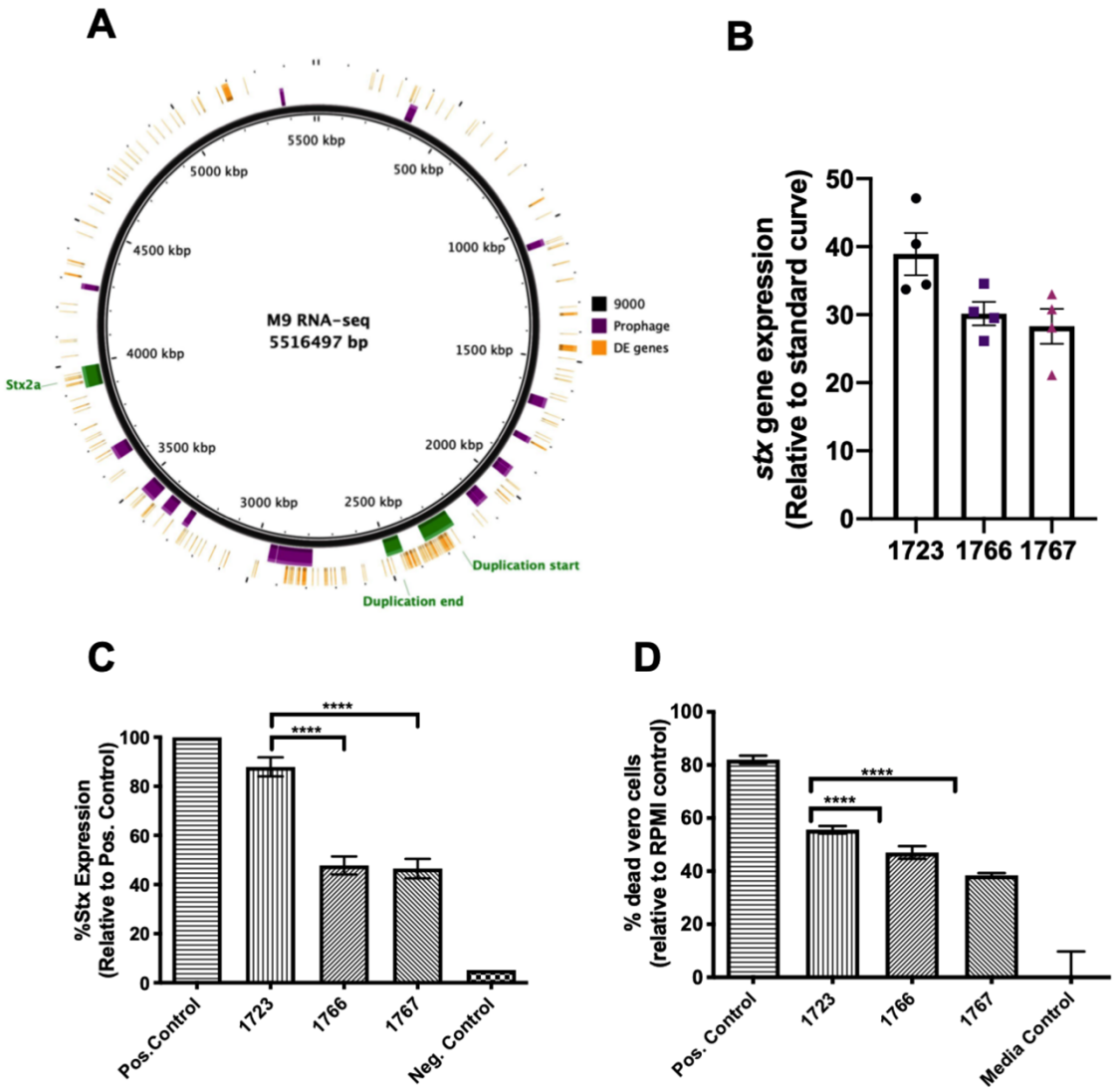

Figure 7.