Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic–uremic syndrome in Germany. N Engl J Med 2011;365:709-17. DOI: 10.1056/NEJMoa1106920.

Table of Contents

I.	Supplementary Methods	2
II.	Supplementary References	12
III.	Supplementary Figure 1	14
IV.	Supplementary Figure 2.	15
V.	Supplementary Figure 3.	16
VI.	Supplementary Figure 4.	17
VII.	Supplementary Figure 5.	18
VIII.	Supplementary Figure 6.	19
IX.	Supplementary Figure 7.	20
X.	Supplementary Figure 8.	21
XI.	Supplementary Figure 9.	28
XII.	Supplementary Figure 10.	29
XIII.	Supplementary Figure 11.	30
XIV.	Supplementary Table 1.	31
XV.	Supplementary Table 2.	32
XVI.	Supplementary Table 3.	33
XVII.	Supplementary Table 4.	34
XVIII.	Supplementary Table 5.	35
XIX.	Supplementary Table 6.	36
XX.	Supplementary Table 7.	38
XXI.	Supplementary Table 8.	40

Supplementary Methods

E. coli Samples

The outbreak strain sequenced in this study is an O104:H4 serotype (referred to as C227-11 in the main text) isolated from a 64-year-old German woman from the City of Hamburg who was hospitalized in the Department of Gastroenterology at Hvidovre University Hospital, Denmark (2011-05-18). Identification of somatic (O) and flagella (H) antigens was carried out by tube and microtitre-plate agglutination with the specific reference sera O1–O181, supplemented with the presumptive new O groups OX182–OX186, and H1–H56[1].]] This patient was hospitalized for diarrhea for less than 24 hours and did not develop HUS. Consumption of salad in Hamburg was suspected to be the source of the infection. C227-11 is an isolate of the pathogen collected independently from other outbreak-associated *E. coli* O104:H4 isolates sequenced by other groups: 1) *E. coli* TY-2482 sequenced by BGI (AFOG01000000); 2) *E. coli* LB226692 sequenced by Life Technologies and collaborators at German National Consulting Laboratory for Hemolytic Uremic Syndrome (HUS) at the Institute of Hygiene, University Hospital Muenster (http://www.ncbi.nlm.nih.gov/nuccore/AFOB00000000); and 3) *E. coli* H112180280 sequenced by the Health Protection Agency of the UK (http://www.hpa.org.uk/).

In addition to the outbreak-associated isolate, we sequenced the genomes of seven *E. coli* O104:H4 isolates from a collection of EAEC strains from Africa and characterized by Statens Serum Institute in Copenhagen (C777-09, C682-09, C35-10, C734-09, C754-09 and C760-09), and four EAEC prototype strains. None of the six recent African isolates had previously been sequenced (see Table 1 in the main text). To examine the distribution of the EAEC isolates relative to EAEC reference isolates, five EAEC prototype strains were included in the genomic comparisons: EAEC strain 042 (O44:H18) was isolated from a child with diarrhea in the course

of an epidemiologic study in Lima, Peru [2]; EAEC strain 17-2 (O3:K-:H2) was isolated from a child with diarrhea in Chile [3]; EAEC strain JM221 (O92:K-:H33) was isolated from an adult with diarrhea in Mexico [4]; EAEC strain C1010-00 (Orough:H-) was isolated from a child with diarrhea in Denmark [5]; and EAEC strain 55989 (O104:H4) was isolated in the course of an etiological study of Human immunodeficiency virus (HIV) and diarrhea in adults in Bangui, Central African Republic [6]. Two of these strains, 55989 [7] and O42 [8] had whole genome sequencing data available prior to our study.

DNA extraction and preparation

For isolate C227-11, we isolated DNA using a Qiagen DNEasy Blood and Tissue Kit per manufacturer instructions. The isolate was grown overnight in standard LB broth and the extraction was performed according to the kit instructions using 1 ml overnight culture per reaction and treating with proteinase K for 2 hours. The DNA was eluted in 100 ul AE buffer per column.

For the seven non-outbreak O104:H4 strains from Africa and four reference isolates, we grew bacterial cultures overnight from a population in 50 ml of Luria broth, minimizing the number of passages of each strain. Genomic DNA was isolated according to standard methods[9]. Briefly, bacterial cells were concentrated by centrifugation, washed and suspended in isolation buffer (0.15 M Tris, 0.1 M EDTA, pH 8.0). Sodium dodecyl sulfate was then added to 1% (vol/vol) final concentration and allowed to incubate for 1 hour at 55ºC or until the solution cleared. Two volumes of phenol:chloroform:isoamylalcohol (25:24:1) were added and briefly mixed by vortex. The resulting solution was separated by centrifugation at 12,000xg for 15 minutes at 4ºC. The aqueous layer (top) was removed to a new tube and mixed with two volumes

of chloroform. The mixture was separated by centrifugation at 12,000xg for 15 minutes at 4ºC and the aqueous layer (top) was moved to a clean tube. The aqueous layer was then extracted with at least 10 volumes of ice-cold ethanol and the precipitated DNA was spooled out of the mixture and suspended in ultrapure water. The purified mixture was further digested with RNase at 37ºC and re-precipitated with 0.1 volumes of 3M NaOAc and 10 volumes of ice-cold ethanol. The pellet was allowed to air dry and then dissolved in a minimal volume of nuclease-free water (Ambion). The quantity and quality of genomic DNA was verified by gel electrophoresis, spectroscopy and picogreen assay.

DNA Sequencing

For all samples, genomic DNA was sheared to a target size 8-10 kb using a Hydroshear Plus Shearing system (Digilab). The outbreak strain C227-11 was also sheared to target size 700bp using a Covaris S200 sonicator (Covaris). Sheared samples were purified, size-selected, and concentrated using Ampure XP Solid Phase Reversible Immobilization (Beckman Coulter) at 0.45-fold volume (8-10Kb) or 0.7-fold volume (700bp) [10]. SMRTBell libraries were constructed with Pacific Biosciences' commercial Template Prep Kit and accompanying protocols [11]. SMRTbell templates were complexed to polymerase molecules using 6nM of the respective SMRTbell library and 3X excess DNA polymerase at a concentration of 18nMas previously described[10,12].

Large insert (8K-10K) SMRTbell libraries were immobilized at 500 pM for 30 minutes on nanofabricated chips containing an array of sequencing zero-mode waveguides (ZMWs). Smaller insert, 700bp SMRTbell complexes were immobilized at 100-200 pM concentrationStandard sequencing was conducted on a PacBio-*RS* Sequencer using 75 or 90

minute continuous collection times[10,12]. The 75-90 minute sequencing period enabled the collection of sequences as long as 15,000bp, where 8kb SMRTbells generated long sub-read length sequences (Supplementary Table 1) and 90 minute collection spans using 700bp SMRTbells provided enough SMRTbell passes for higher accuracy CCS sequencing.

De novo genome assembly of SMRT sequencing data

The *de novo* assembly algorithm applied to the SMRT sequencing data generated on the C227-11 isolate incorporates elements of the AMOS assembly software package[13] and employs several novel algorithms tailored to Pacific Biosciences' continuous long read (CLR) and circular consensus sequencing (CCS) read data. Supplementary figure 1 provides a high-level overview of the assembly pipeline. In the first phase, PacBio CCS data were aligned to CLR data using BLASR (http://www.pacbiodevnet.com/) to allow error correction. Error-corrected sequence for each long-read was derived by performing a multiple-sequence alignment of CCS reads followed by plurality consensus base-calling using the AMOS algorithm make-consensus. The CLR data was split at any point without CCS coverage.

The error corrected CLR data were then fed directly into the Pacific Biosciences assembler ALLORA (http://www.pacbiodevnet.com/) (for A Long Read Assembler). ALLORA uses a traditional overlap-layout-consensus approach. Pairwise overlap alignments were detected using BLASR and analyzed to identify maximally contiguous sequences (contigs) using the AMOS algorithm tigger. Consensus sequence for each contig was again derived using the AMOS algorithm make-consensus. As a final step, minimus2 was run on the resultant contigs to merge any remaining overlaps between contigs. The final contigs were used as a reference to the

resequencing pipeline, and the entirety of CCS data was aligned against them in order to provide the final consensus sequence.

Resequencing analysis

The complete genome sequences for isolates TY-2482 and 55989 were used as reference genomes for mapping all sequencing reads generated on the 12 strains and isolates indicated above. Mapping was carried out using the FASTA sequences found at http://climb.genomics.cn/Ecoli_TY-2482 (doi:10.5524/100001 for TY-2482) and NCBI accession numbers NC_011748 (for 55989). The re-sequencing analysis pipeline consisted of filtering, reference alignment, and consensus calling steps. Reads were filtered by requiring that each raw read had a raw read length > 100bp and an estimated accuracy of at least 75%. This reduced the number of reads from non-sequencing zero-mode waveguides while removing few true sequencing reads since the raw readlength averaged ~2000bp in these data. Reads were aligned to the TY-2482 and 55989 references using a computational variant of the Smith-Waterman algorithm, which finds chains of exact sequence hits in the reference sequence and refines these chains into Smith-Waterman local alignments. A multiple sequence alignment (MSA) was constructed by using the reference-guided pairwise alignments. Consensus sequences were called for each isolate using a majority-rule plurality algorithm.

Supplementary Table 1 summarizes the raw statistics of the sequencing reads. The single pass mean accuracy achieved over all samples for the CLR data was 84.4%. However, the accuracy distribution is asymmetric (Supplementary Figure 2), with a heavy left tail that is longer than realized previously[10], due to the fact that very long reads (CLR data) can be mapped with extremely high precision even at low accuracy. Therefore, one consequence of longer reads

is a reduced mean-level accuracy given long, low accuracy reads map perfectly and aid in scaffolding higher accuracy short reads for de novo assembly (so filtering out such reads would result in less coverage and, as a result, a less robust de novo assembly). Given the heavy tail on the left hand side of this distribution, a more accurate representation of the accuracy distribution is the mode (Supplementary Table 1; Supplementary Figure 11), which on average for the CLR data was 88.2%. The circular consensus sequence accuracy for the CCS data generated on isolate C227-11 for de novo assembly had a mean level accuracy of 97.8%, but an accuracy characterized by the mode of this distribution of 99.9%. The combination of highly accurate CCS data combined with the very long CLR read data resulted in a high-quality assembly of the C227-11 genome comprised of 37 contigs covering the chromosome, with 30 of the largest contigs covering 99% of the genome and all 37 contigs combined covering 99.7% of the genome (Supplementary Table 3).

To investigate the effect of the filtering step on the data, we also aligned reads to the reference without imposing any pre-alignment filtering. Only significant hits were kept to the reference as measured by comparison to a random alignment model. Under these conditions, average read length of the underlying sequence was 2,110bp and the sequenced depth of coverage was increased by 44-66%, without an appreciable distortion in the coverage uniformity, demonstrating that the filtering process was unlikely to introduce errors into downstream sequence analyses. These results suggest that for the purposes of coverage analysis, we should be able to derive more power from the same input data by allowing more reads through the early analysis stages and applying significance tests after reference alignment.

Sequence Coverage Analysis

The depth of coverage across the TY-2482 and 55989 genome sequences were plotted for each of the O104 isolates we sequenced. For Figure 1 in the main text, the coverage variation was estimated using a breakpoint analysis, given breakpoints are easy to detect with the CLR data. For each position the mean of the minimum of the left and right fragment lengths for a mapped read is computed. If there is no breakpoint this estimate will on average be equal to one half the mean mapped subread length at that position. If a breakpoint exists at a given position, this mean fragment length estimate will be significantly smaller than the expected mean mapped subread length. Colors were assigned across the continuum of possible coverage estimates, with estimates close to the expected average taking on cooler colors (blue) and those taking on values significantly less than the expected average taking on hotter colors (red). The mean values and corresponding standard deviations of the minimum of the left and right fragment lengths were 752±103 bp, 804 ±264 bp, 873 ± 222 bp, 728 ± 199bp, 767 ±207 bp, 767 ± 265 bp, 773 ± 307 bp and 730±355 bp for strains 55989, C227-11, C734-09, C35-10, C682-09, C760-09, and C754-09, and C777-09 respectively.

Genome annotation of C227-11

Assembly contigs for C227-11 were analyzed using a previously described frameshift prediction procedure [14]. This procedures locates apparent frameshifts in gene coding regions by incorporating them into a hidden Markov model (HMM) trained on the prokaryotic gene coding model. Frameshifts in sequenced DNA can be indicative of uncorrected indel sequencing errors, and their presence pinpoints areas in which polishing of the DNA sequence is required, given uncorrected errors of this type can result in interrupted gene models and so are disruptive to downstream analyses. In addition, given a significant indel rate in the final assembly

(Supplementary Table 2), we carried out an extra step of verifying apparent frameshifts against the Pfam protein database. Regions containing such verified apparent frameshifts were interrogated for presence of indel sequencing errors, and automatic error correction was performed where errors were confirmed. Of 285 regions with verified apparent frameshifts, 175 were successfully corrected using this scheme.These changes are not included when computing the identity/accuracy of the *de novo* C227-11 assembly as reported in the main text and Supplementary Table 3.

After polishing the genome sequnce for C227-11, automated genome annotation was carried out using a previously described computational genomics pipeline [15]. In this pipeline, genes are predicted using a combination of the Glimmer and GeneMarkS de novo gene predictors, and BLASTP alignment against the nonredundant nucleotide database (nt). Annotation was performed using the InterProScan family of programs, the SignalP algorithm and comparison against the VFDB virulence factor database. The resulting annotation, formatted according to standard terminology, was submitted to GenBank with the genome sequence under GenBank accession number AFST00000000.

Comparison of genome assemblies for German outbreak isolates

Supplemental Table 2 and Supplementary Figure 3 summarize a comparison of each of the publicly available datasets to the completed TY-2482 reference. The Mummer package[16] was used to align the genomes, carry out SNP calling, and as a basis for computing reference coverage and identity for each of the assemblies. Parameters optimized for SNP-calling were used (nucmer with the "–maxmatch" parameter, delta-filter with the "-1" parameter , and show-snps with the "-Clr" parameters). Specifically, these parameters require 1-1 alignment of query

and reference intervals, but allow query sequences to be rearranged in order to fit the reference; they also minimize repeat mappings when calling variants.

The single nucleotide differences identified between the TY-2482 and C227-11 assemblies may reflect actual sequence differences between the isolates or sequencing errors specific to either of the sequences. The PCR amplification required for sequencing the TY-2482 genome by Illumina's HiSeq 2000 instrument and Life Technology's Personal Genome Machine instrument, can introduce errors in the template sequence as well as amplification bias, which in turn can lead to systematic errors in the final assemblies[17-19]. On the other hand, the sequencing of C227-11 using the PacBio RS did not require PCR amplification as a step in the sequencing process, so that when combined with the uniform nature of errors on the PacBio RS platform (lack of systematic bias), it is not unreasonable to expect that the consensus accuracy could be higher even despite a high single pass sequencing error rate[11]. By combining these data generated from different platforms we have reduced the overall rate of error in our finished genome sequence for C227-11 (last row of Supplementary Table 2).

A small number of larger scale structural rearrangements were also observed between C227-11 and TY-2482 (Supplementary Figure 5). Distinguishing misassembles from true variation is often challenging. To validate large rearrangements two methods were taken: 1) comparison to other strains (notably H1121), and 2) remapping of long read data to examine if the alternative (TY-2482) could be accepted. Supplementary Figure 5 shows two examples of larger events shared between C227-11 and H1121. Long-read data was remapped to the TY-2482 reference versus the H1121/C227-11 assemblies, showing strong support of the H1121/C227-11 preferentially over the TY-2482 assemblies in this region.

Whole genome phylogenetic analysis

The sequence data for 40 *E. coli/Shigella* genomes (Supplementary Table 3) was downloaded from GenBank and combined with sequence data from the 12 EAEC isolates. Sequences were aligned with Mugsy [20], which incorporates MUMmer [21,22] and SeqAn [23] to generate blocks of conserved, aligned sequence between species in the MAF file format. Blocks were then joined together and converted to a multifasta file with the bx-python toolkit (http://bitbucket.org/james_taylor/bx-python/wiki/Home). Columns with gaps in any one genome were removed with Gblocks [24] to create the core alignment, which consists of ~2.56 Mb of genomic sequence. As a result of this process, the indels highlighted in Supplementary Table 2 and Supplementary Figure 3 were not considered in the phylogenetic analysis. A phylogenetic tree was inferred by RAxML[25] with one hundred bootstrap replicates and a general time-reversible model. Additionally, a subtree containing only EAEC isolates were compared in the same manner.

Quantitative Reverse Transcriptase Polymerase Chain Reaction (qRT-PCR) to examine the induction of the Shiga-toxin genes by ciprofloxacin.

An overnight culture of *E. coli* strain C227-11 was diluted in Lauria broth to an OD_{600} nm of 0.09 and divided into six separate cultures of equal volume. Ciprofloxacin at a final concentration of 25 ng/mL was added to three of the cultures. The C227-11 cultures were incubated at 37°C for 6 hrs. and RNA was extracted using the Ambion RiboPure-Bacteria Kit, DNaseI treatment (Ambion). The resulting RNA was assayed using qRT-PCR with technical duplicates of the triplicate biological samples. The primers for the *stx2b* and *rpoA* genes were previously described by Zhang et al [26] and Rasko et al [27], respectively. The qRT-PCR was

preformed and analyzed as previously described in Rasko et al. [27] and reported as fold change as fold change of the *stx2b* gene, using the *rpoA* gene as a control.

Identification of E. coli virulence factors

Sequences were obtained from GenBank for known and characterized *E. coli* virulence factors from the EAEC, EPEC and EHEC pathovars. A list of the examined genes can be found in Supplementary Table 4.

Nucleotide sequence accession numbers

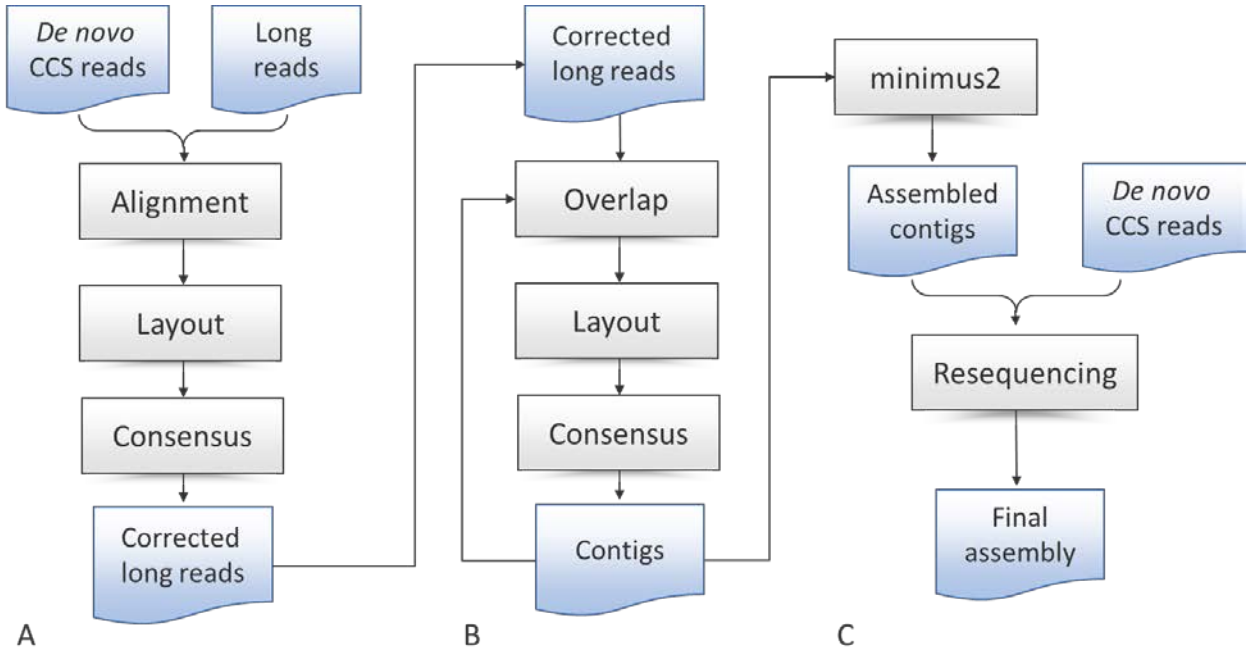
The genomic data for these isolates has been submitted to GenBank under accession number AFST00000000 (C227-11) and to the NCBI Sequence Read Archive under study number SRA038239.1 (all strains).

Supplementary References

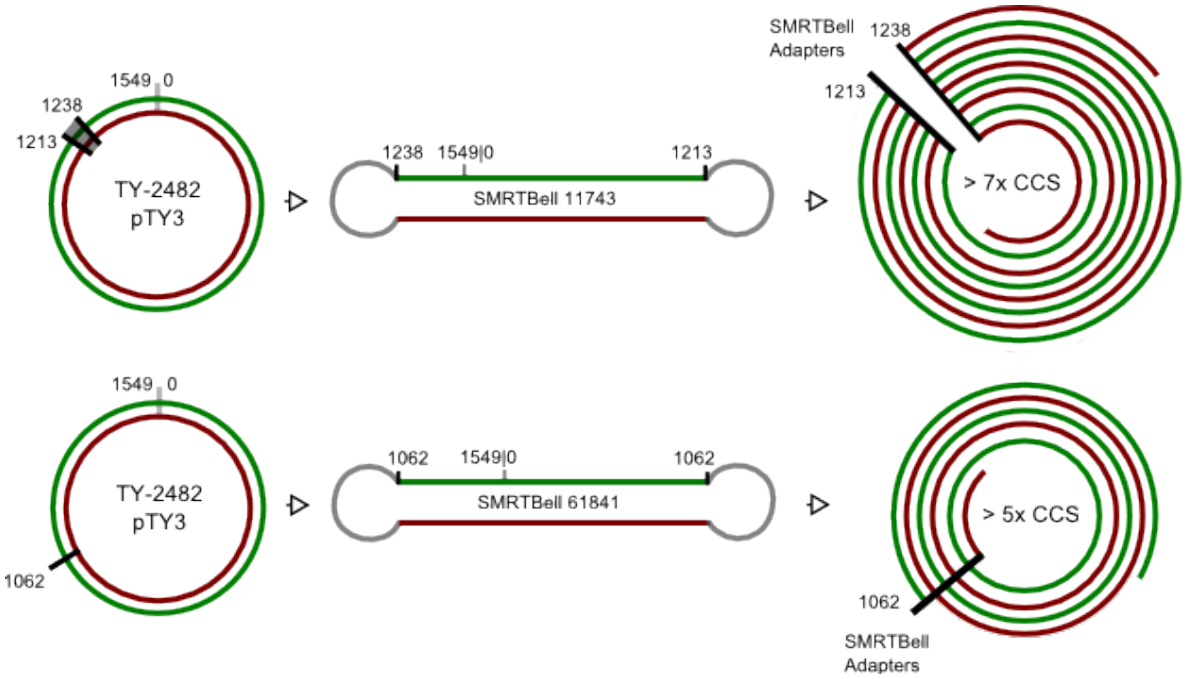
1. Orskov F, Orskov I. Escherichia coli serotyping and disease in man and animals. Can J Microbiol 1992;38:699-704.
2. Nataro JP, Baldini MM, Kaper JB, Black RE, Bravo N, Levine MM. Detection of an adherence factor of enteropathogenic Escherichia coli with a DNA probe. J Infect Dis 1985;152:560-5.
3. Vial PA, Robins-Browne R, Lior H, et al. Characterization of enteroadherent-aggregative Escherichia coli, a putative agent of diarrheal disease. J Infect Dis 1988;158:70-9.
4. Mathewson JJ, Oberhelman RA, Dupont HL, Javier de la Cabada F, Garibay EV. Enteroadherent Escherichia coli as a cause of diarrhea among children in Mexico. J Clin Microbiol 1987;25:1917-9.
5. Olesen B, Neimann J, Bottiger B, et al. Etiology of diarrhea in young children in Denmark: a case-control study. J Clin Microbiol 2005;43:3636-41.
6. Germani Y, Minssart P, Vohito M, et al. Etiologies of acute, persistent, and dysenteric diarrheas in adults in Bangui, Central African Republic, in relation to human immunodeficiency virus serostatus. Am J Trop Med Hyg 1998;59:1008-14.

7. Touchon M, Hoede C, Tenaillon O, et al. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 2009;5:e1000344.
8. Chaudhuri RR, Sebaihia M, Hobman JL, et al. Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative Escherichia coli strain 042. PLoS One 2010;5:e8801.
9. Ge Z, Taylor DE. *H. pylori* DNA transformation by natural competence and electroporation. In: Clayton CL, Mobley HLT, eds. *Helicobacter pylori* Protocols. Totowa: Humana Press Inc.; 1992:145-52.
10. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med 2011;364:33-42.
11. Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res;38:e159.
12. Korlach J, Bjornson KP, Chaudhuri BP, et al. Real-time DNA sequencing from single polymerase molecules. Methods Enzymol 2010;472:431-55.
13. Pop M, Phillippy A, Delcher AL, Salzberg SL. Comparative genome assembly. Brief Bioinform 2004;5:237-48.
14. Antonov I, Borodovsky M. Genetack: frameshift identification in protein-coding sequences by the Viterbi algorithm. J Bioinform Comput Biol 2010;8:535-51.
15. Kislyuk AO, Katz LS, Agrawal S, et al. A computational genomics pipeline for prokaryotic sequencing projects. Bioinformatics 2010;26:1819-26.
16. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. Genome Biol 2004;5:R12.
17. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31-46.
18. Schatz MC, Delcher AL, Salzberg SL. Assembly of large genomes using second-generation sequencing. Genome Res 2010;20:1165-73.
19. Whiteford N, Skelly T, Curtis C, et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics 2009;25:2194-9.
20. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 2011;27:334-42.
21. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 2002;30:2478-83.
22. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. Curr Protoc Bioinformatics 2003;Chapter 10:Unit 10 3.
23. Doring A, Weese D, Rausch T, Reinert K. SeqAn an efficient, generic C++ library for sequence analysis. BMC Bioinformatics 2008;9:11.
24. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564-77.
25. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 2006;22:2688-90.
26. Zhang Y, Laing C, Zhang Z, et al. Lineage and host source are both correlated with levels of Shiga toxin 2 production by Escherichia coli O157:H7 strains. Appl Environ Microbiol 2010;76:474-82.
27. Rasko DA, Moreira CG, Li de R, et al. Targeting QseC signaling and virulence for antibiotic development. Science 2008;321:1078-80.

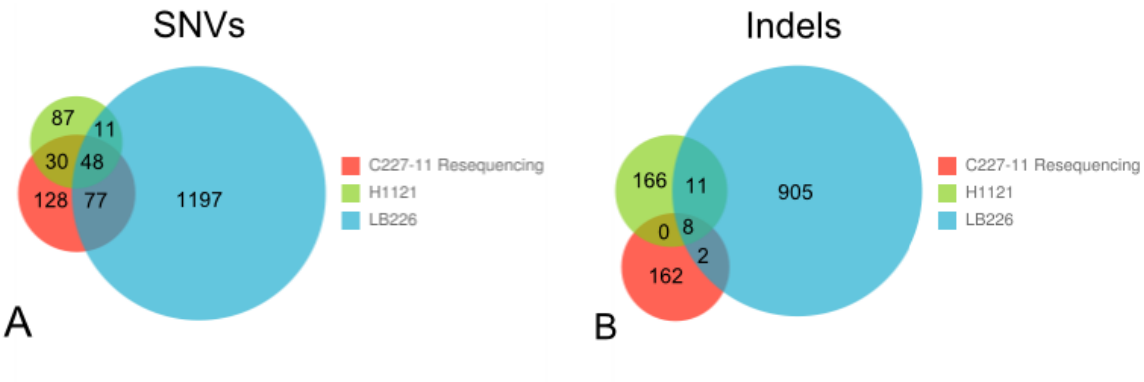
Supplementary Figures



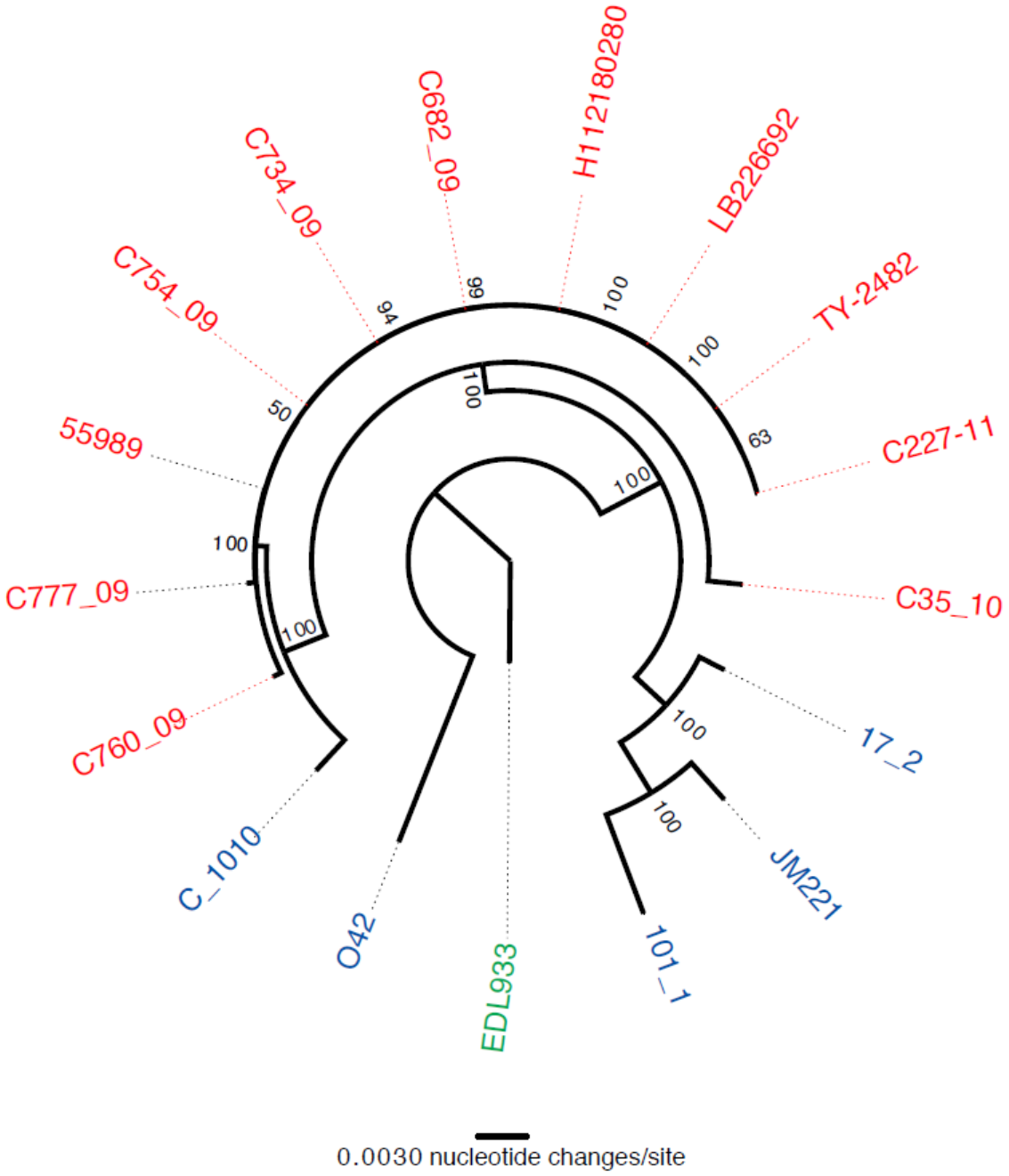
Supplementary Figure 1. *De novo a*ssembly pipeline for SMRT sequencing data. **A)** The first step of the assembly process uses highly accurate CCS reads to correct errors in the single-pass long sequence reads. **B)** The corrected long reads are then provided as input to ALLORA, an iterative overlap-layout-consensus *de novo* assembly algorithm. **C)** The resulting assembled contigs are then polished with the resequencing pipeline and the original CCS reads, producing the final, high-accuracy assembly.



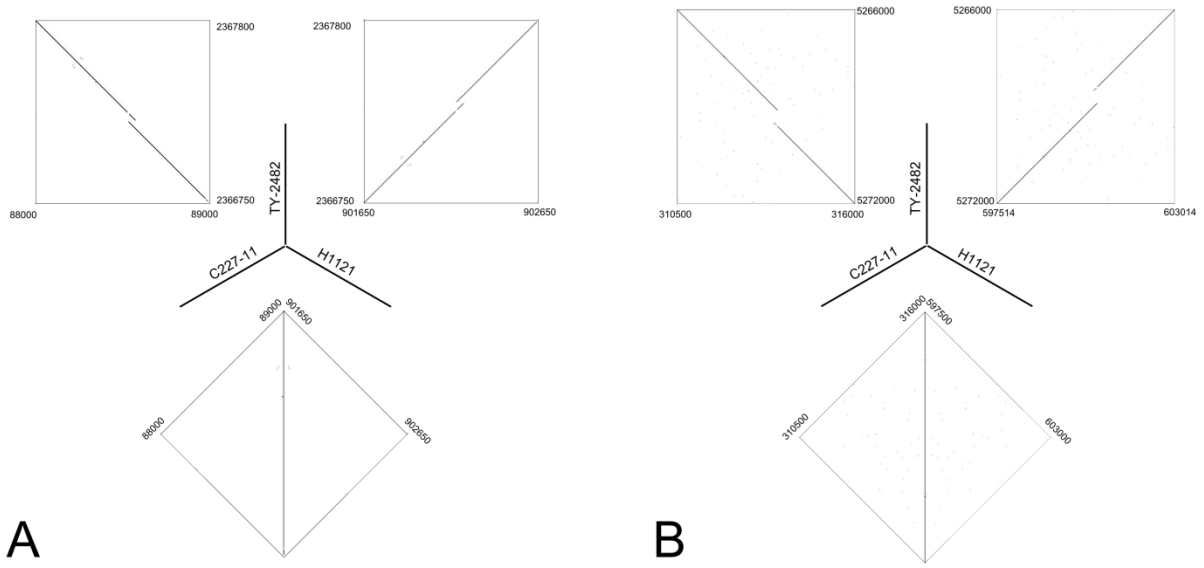
Supplementary Figure 2. Single reads spanning a small plasmid. During the automated assembly process of strain C227-11 a small contig was detected with strong homology to the pTY3 plasmid from the TY-2482 assembly. Further analysis revealed the presence of a number of long reads clearly representing SMRTBell structures containing the majority of the pTY3 sequence. The location of apparent break points within the plasmid varied, suggesting that multiple plasmid sequences were linearized during sample preparation and subsequently ligated to adapter sequences to form SMRTBells. In some cases small portions of the plasmid surrounding the break point were absent from the SMRTBell, possibly due to degradation of the linearized plasmid before ligation. For two reads selected from Supplementary Table 2 a likely model for conversion of the plasmid into a SMRTBell and the resulting sequence read are shown. Shaded regions within the plasmid diagrams for pTY3 depict the region of the plasmid not present in the corresponding read. Single reads (far right) are represented in a circular form with adapters removed for clarity. All coordinates are relative to the TY-2482 pTY3 sequence. Green and red lines indicate forward and reverse orientation.



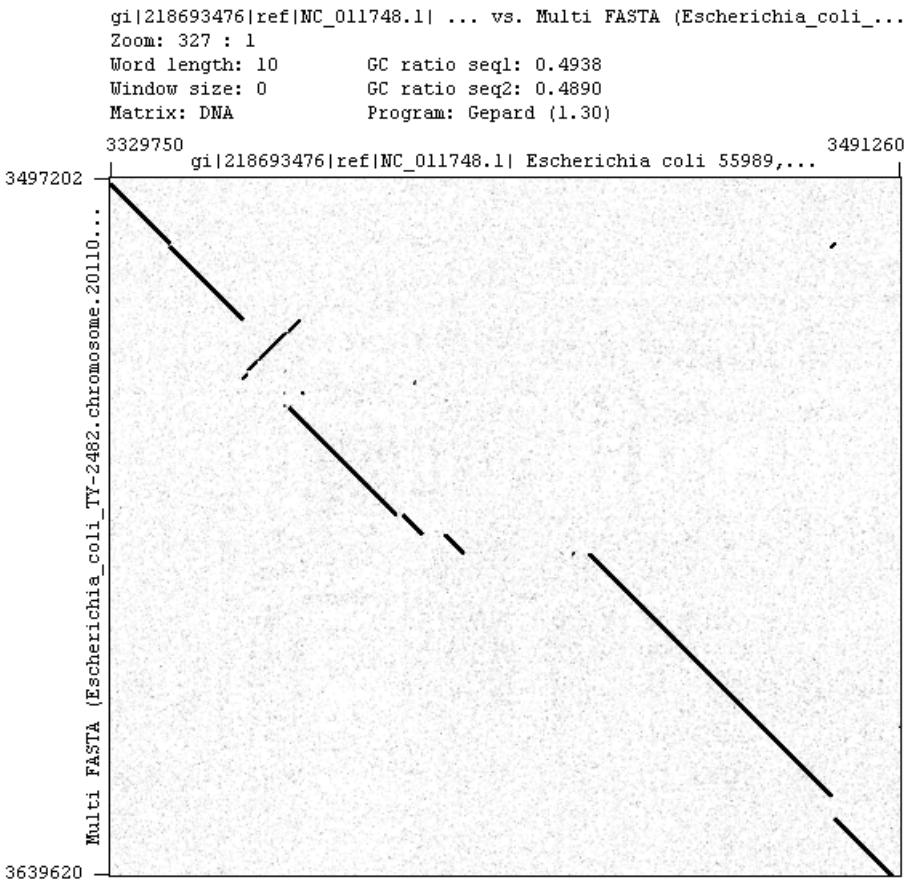
Supplementary Figure 3. Venn Diagram of single nucleotide variation calls relative to the TY-2482 chromsome. The sum of the counts in each colored circle represents the number of single nucleotide variations (Panel A) and single nucleotide insertions or deletions (indels; Panel B) detected between the indicated isolate and TY-2482 (Supplementary Table 3). The sizes of the circles are proportional to the number of variants detected. The counts in the overlap regions indicate whether SNVs or Indels were shared between the indicated isolates. C227-11 variant calls were made using EviCons, accepting calls with consensus confidence scores greater than 1 (EviCons consensus confidences range from 0 to 10). Allowing variant calls with lower confidence increases the number of predicted SNVs to 284 and indels to 882.



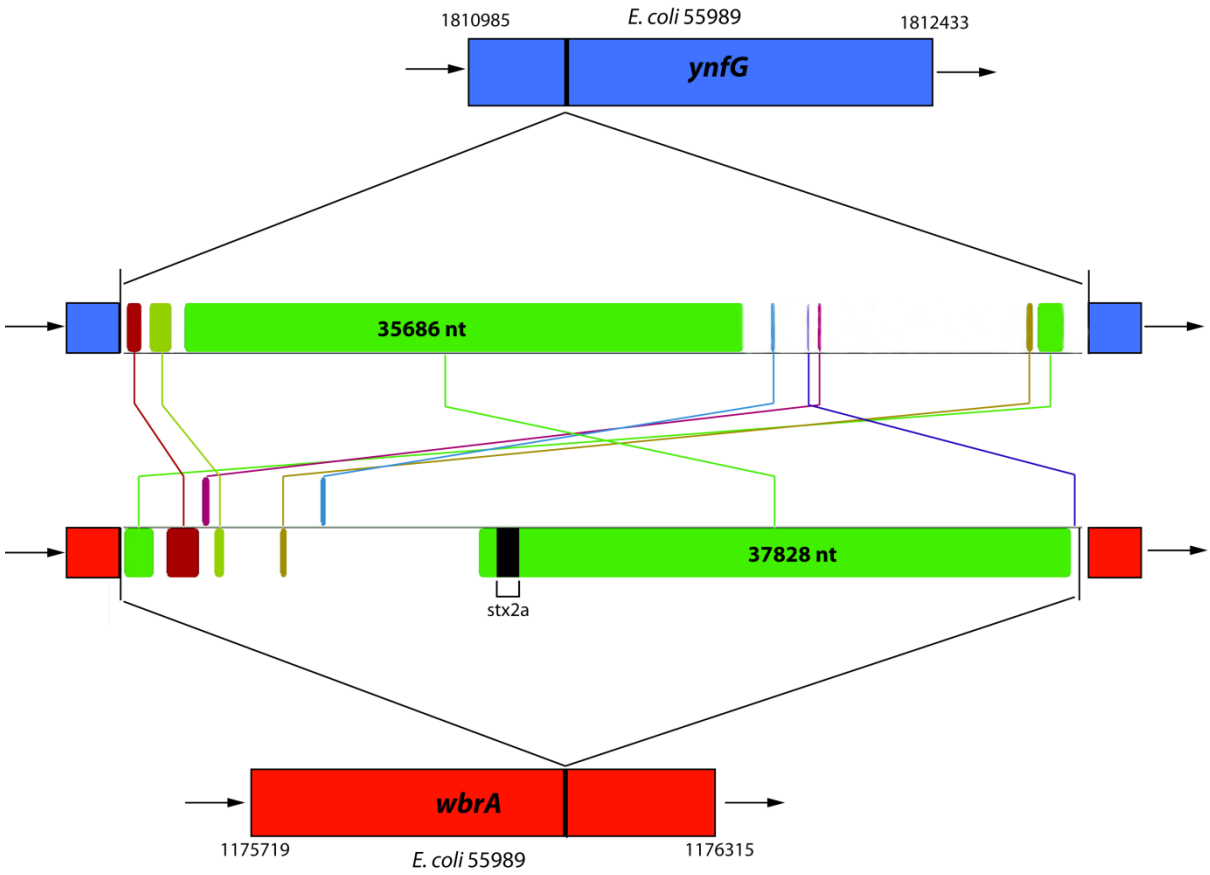
Supplementary Figure 4. Phylogenetic comparisons of *E. coli/Shigella* isolates to determine the EAEC-specific phylogeny. This represents a more detailed view of the phylogeny of the EAEC isolates than depicted and described in Figure 2 of the main text.



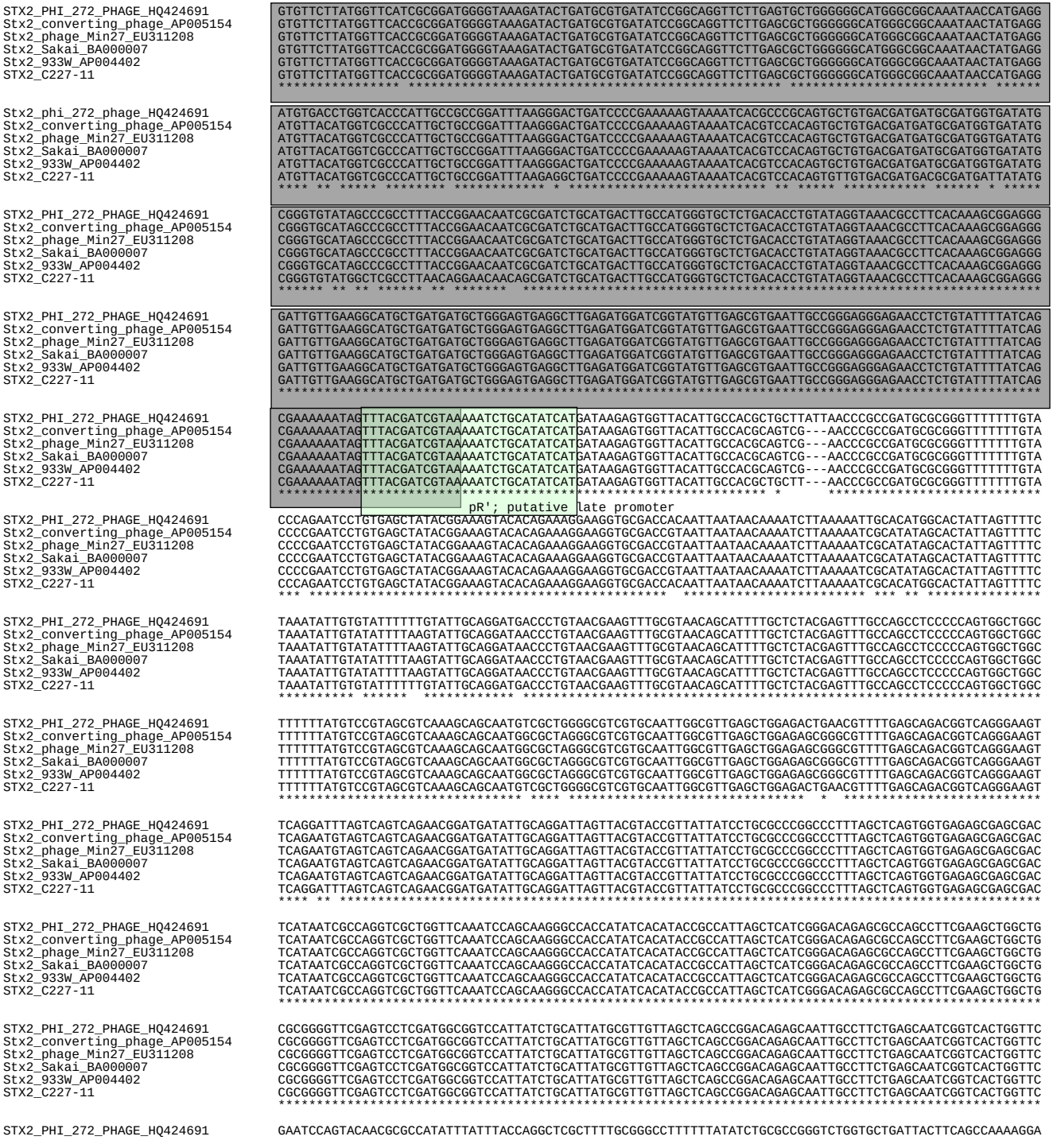
Supplementary Figure 5. Comparison of Large Structural Differences between C227-11, H1121 and TY-2482. In each plot three axes represents regions from TY-2482, C227-11 and H1121. For each pair of axes a dot plot for the corresponding strain pair is plotted. If the genomic regions are identical between the strains, the dot plot gives rise to a continuous diagonal line. If there are large-scale structural variations between the strains, the dot plot will give rise to lines that are disconnected. **A)** The break in the diagonal lines in the comparisons to the TY-2482 chromosome represent a compression of a tandem duplication in C227-11 and H1121 relative to TY-2482. **B)** The break in the diagonal lines in the comparisons to the TY-2482 chromosome represent a deletion in C227-11 and H1121 relative to TY-2482.

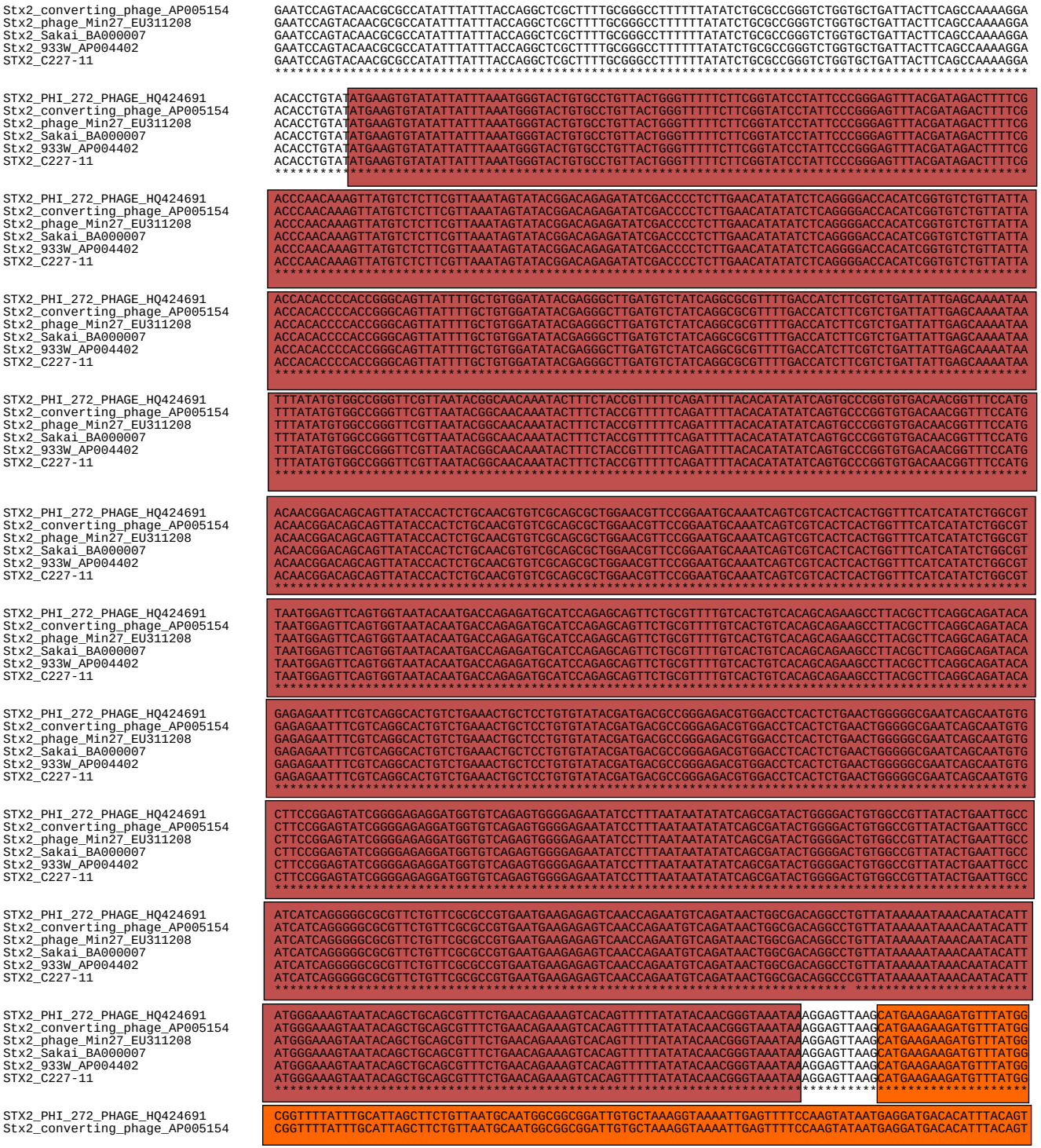


Supplementary Figure 6. Dot plot of the region with significant structural variation between TY-2482 and the 55989 reference EAEC strain. This region corresponds to the green shaded regions highlighted in the chromosome coverage plot in Figure 1 of the main text. This region harbors a number of virulence factors like *pic* and the *aai* pathogenicity island. While these virulence factors are present in both TY-2482 and 55989, there is significant structural variation between these two strains in this region, as depicted in this plot.



Supplementary Figure 7. Similarity between regions of the two lambdoid phages in C227-11. The two phages are present at distinct genomic locations shown with reference to the *E. coli* 55989 genome at the top and bottom of the figure. Although large protions of the phage genomes are homologous, some portions are quite distinct, as demonstrated by the Mauve alignment. Importantly, only the phage inserted into the *wbrA* gene contains the *stx2* genes.


antitermination protein Q
pR'; putative late promoter



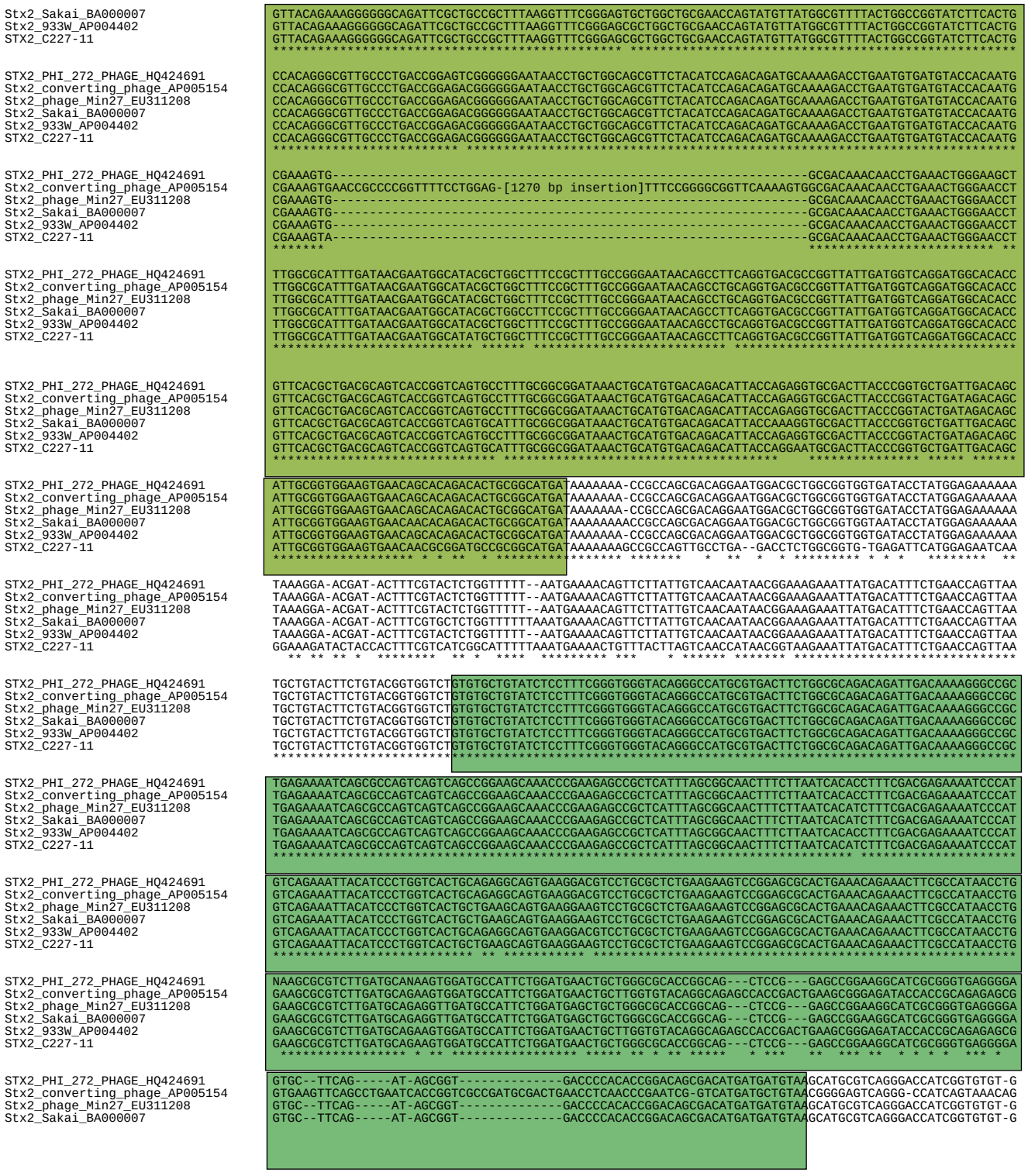
Shiga-toxin 2 subunit A

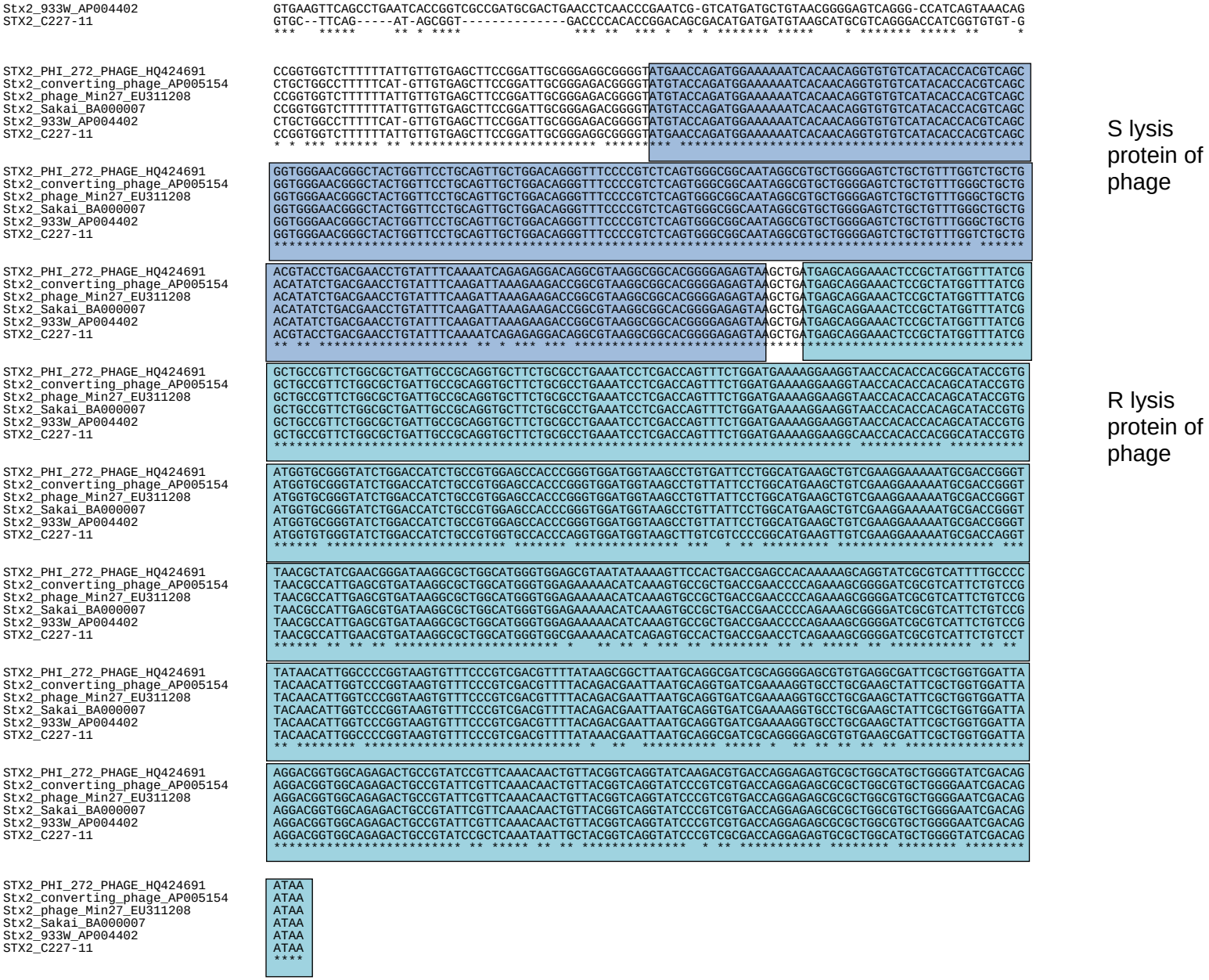
```
Stx2_phage_Min27_EU311208          CGGTTTTATTTGCATTAGCTTCTGTTAATGCAATGGCGGCGGATTGTGCTAAAGGTAAAATTGAGTTTTCCAAGTATAATGAGGATGACACATTTACAGT
Stx2_Sakai_BA000007                CGGTTTTATTTGCATTAGCTTCTGTTAATGCAATGGCGGCGGATTGTGCTAAAGGTAAAATTGAGTTTTCCAAGTATAATGAGGATGACACATTTACAGT
Stx2_933W_AP004402                 CGGTTTTATTTGCATTAGCTTCTGTTAATGCAATGGCGGCGGATTGTGCTAAAGGTAAAATTGAGTTTTCCAAGTATAATGAGGATGACACATTTACAGT
STX2_C227-11                       CGGTTTTATTTGCATTAGCTTCTGTTAATGCAATGGCGGCGGATTGTGCTAAAGGTAAAATTGAGTTTTCCAAGTATAATGAGGATGACACATTTACAGT
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
Stx2_converting_phage_AP005154     GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
Stx2_phage_Min27_EU311208          GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
Stx2_Sakai_BA000007                GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
Stx2_933W_AP004402                 GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
STX2_C227-11                       GAAGGTTGACGGGAAAGAATACTGGACCAGTCGCTGGAATCTGCAACCGTTACTGCAAAGTGCTCAGTTGACAGGAATGACTGTCACAATCAAATCCAGT
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
Stx2_converting_phage_AP005154     ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
Stx2_phage_Min27_EU311208          ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
Stx2_Sakai_BA000007                ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
Stx2_933W_AP004402                 ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
STX2_C227-11                       ACCTGTGAATCAGGCTCCGGATTTGCTGAAGTGCAGTTTAATAATGACTGAGGCATAACCTGATTCGTGGTATGTGGGTAACAAGTGTAATCTGTGTCAC
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
Stx2_converting_phage_AP005154     AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
Stx2_phage_Min27_EU311208          AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
Stx2_Sakai_BA000007                AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
Stx2_933W_AP004402                 AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
STX2_C227-11                       AATTCAGTCAGTTGACAGTTGCCTGTCAGACTGAGCATTTGTTAAAAAAATTTCGCATGGTGAATCCCCCTGTGTGGAGGGGCGACTGGTGAAAAATCCT
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
Stx2_converting_phage_AP005154     TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
Stx2_phage_Min27_EU311208          TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
Stx2_Sakai_BA000007                TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
Stx2_933W_AP004402                 TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
STX2_C227-11                       TGCTTGTGATTCATTATCGACACGGGTTCGGTGGTACCAGGCCGAACTCACCGGGAGGCACCCGGCACCATGCAGTATACAGAGATTAGGCATATACCAA
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAATCATTTTAACCA
Stx2_converting_phage_AP005154     GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAATCATTTTAACCA
Stx2_phage_Min27_EU311208          GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAATCATTTTAACCA
Stx2_Sakai_BA000007                GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAATCATTTTAACCA
Stx2_933W_AP004402                 GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAATCATTTTAACCA
STX2_C227-11                       GGCCTCTCATAGCAGGGGCCTTTTTACATGTAAAAAAGCCCGAGTGGGTTCGGGCAATTGCATGAGATACTCGTTTTAATAATCGAAAGCATTTTAACCA
                                   **************************************************************************************** ***********

STX2_PHI_272_PHAGE_HQ424691        GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
Stx2_converting_phage_AP005154     GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
Stx2_phage_Min27_EU311208          GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
Stx2_Sakai_BA000007                GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
Stx2_933W_AP004402                 GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
STX2_C227-11                       GGATTCATAAGGCTGCGCAACTGCGCGGCCTTTTTCGTATTTCGGGCTGTAGTCTTCCTTCTGCCATTGTCCTGTAACTTCCGGACTTCAGCCCGCTCCT
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        CATCTGACTCACA-CATTATCCCGACCGGGAGGATTCATGGCATTTAAACACTATGACGTGGTCAGGGCGGCGTCGCCGTCAGACCTTGCGGAACGAATA
Stx2_converting_phage_AP005154     TATTTACTCACAATATTATCCCGGCCGGGAGGATTCATGGCATTTAAACACTATGATGTTGTCAGGGCGGCGTCGCCGTCAGACCTTGCGAAACGAATA
Stx2_phage_Min27_EU311208          TATTTACTCACAATATTATCCCGGCCGGGAGGATTCATGGCATTTAAACACTATGATGTTGTCAGGGCGGCGTCGCCGTCAGACCTTGCGAAACGAATA
Stx2_Sakai_BA000007                TATTTACTCACAATATTATCCCGGCCGGGAGGATTCATGGCATTTAAACACTATGATGTTGTCAGGGCGGCGTCGCCGTCAGACCTTGCGAAACGAATA
Stx2_933W_AP004402                 TATTTACTCACAATATTATCCCGGCCGGGAGGATTCATGGCATTTAAACACTATGATGTTGTCAGGGCGGCGTCGCCGTCAGACCTTGCGAAACGAATA
STX2_C227-11                       CATCTGACTCACA-CATTATCCCGACCGGGAGGATTCATGGCATTTAAACACTATGACGTGGTCAGGGCGGCGTCGCCGTCAGACCTTGCGGAACGAATA
                                    ** * *******  ********* ******************************** ** ****************************** ********

STX2_PHI_272_PHAGE_HQ424691        ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
Stx2_converting_phage_AP005154     ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
Stx2_phage_Min27_EU311208          ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
Stx2_Sakai_BA000007                ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
Stx2_933W_AP004402                 ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
STX2_C227-11                       ACTCAAAAACTGAAGGAAGGGTGGCAGCCTTATGGTAGTGCGCTGATTTCGACAGCTGGTTATGGTGCGGAGTTCATCCAGCCAGTTGTGAGTGAGGGGA
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
Stx2_converting_phage_AP005154     GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
Stx2_phage_Min27_EU311208          GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
Stx2_Sakai_BA000007                GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
Stx2_933W_AP004402                 GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
STX2_C227-11                       GCATCTCATCACCAGAGGAGCCAGGCAACCGTCCGACGACCTCAGCGCCTTCTGTTGCGCCAGAATATTACTATGTGATCGCGCTTGCTGGTCAGTCCAA
                                   ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691        TGGTATGTCATACGGTGAGGGACTGCCATTACCGGATACATTCGACAGTCCTGATCCACGAATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGA
Stx2_converting_phage_AP005154     TGGTATGTCATACGGTGAGGGACTGCCATTGCCGGATACATTCGACAGCCCTGATCCACGTATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGC
Stx2_phage_Min27_EU311208          TGGTATGTCATACGGTGAGGGACTGCCATTGCCGGATACATTCGACAGCCCTGATCCACGTATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGC
Stx2_Sakai_BA000007                TGGTATGTCATACGGTGAGGGACTGCCATTGCCGGATACATTCGACAGCCCTGATCCACGTATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGC
Stx2_933W_AP004402                 TGGTATGTCATACGGTGAGGGACTGCCATTGCCGGATACATTCGACAGCCCTGATCCACGTATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGC
STX2_C227-11                       TGGTATGTCATACGGTGAGGGACTGCCATTACCGGATACATTCGACAGTCCTGATCCACGAATTAAACAGTTAGCGCGTCGCAGTACGGTGACACCGGGA
                                   ****************************** ***************** *********** **************************************

STX2_PHI_272_PHAGE_HQ424691        GGTGCCGCCTGTAAGTATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGCCTTAACCATCCGAAAGCGGACCTGTCAA
Stx2_converting_phage_AP005154     GGTGCAGTATGCAAATATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGTCTTAACCATCCGAAAGCGGACCTGTCAA
```

Shiga-toxin 2 subunit B

Hypothetical protein

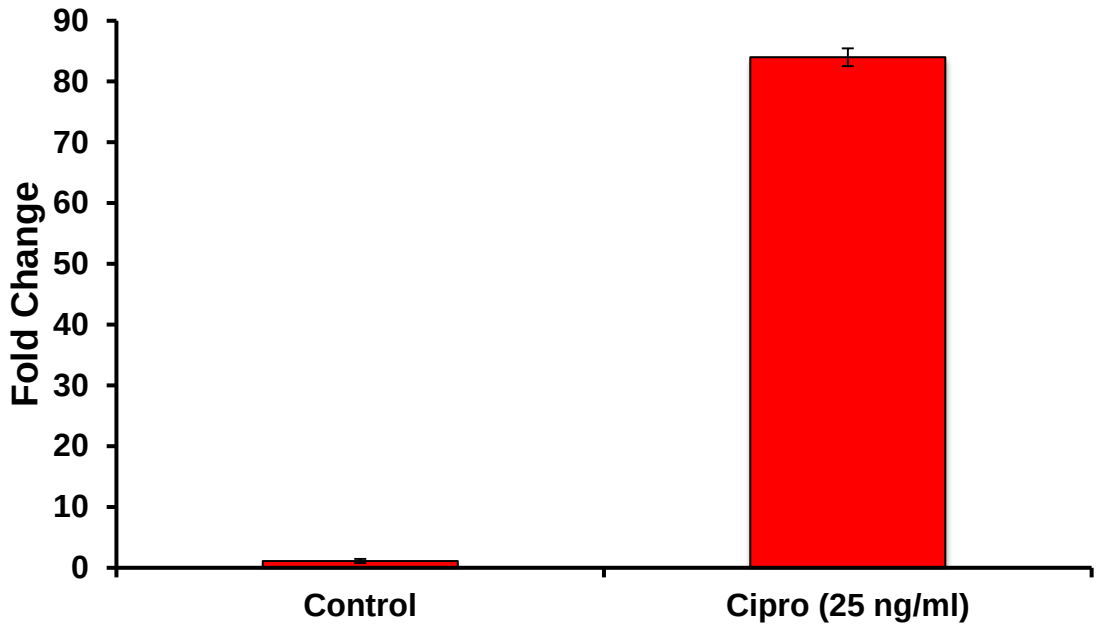
```
Stx2_phage_Min27_EU311208           GGTGCAGTATGCAAATATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGTCTTAACCATCCGAAAGCGGACCTGTCAA
Stx2_Sakai_BA000007                 GGTGCAGTATGCAAATATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGTCTTAACCATCCGAAAGCGGACCTGTCAA
Stx2_933W_AP004402                  GGTGCAGTATGCAAATATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGTCTTAACCATCCGAAAGCGGACCTGTCAA
STX2_C227-11                        GGTGCCGCCTGTAAGTATAACGACATCATTCCGGCGGACCATTGTCTGCATGATGTGCAGGACATGAGCCGCCTTAACCATCCGAAAGCGGACCTGTCAA
                                    ***** *  ** ** ********************************************************* ****************************

STX2_PHI_272_PHAGE_HQ424691         AGGGGCAGTACGGAACCGTGGGGCAGGGGCTGCATATCGCCAAAAAATTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
Stx2_converting_phage_AP005154      AGGGGCAGTACGGAACCGTGGGGCAGGGGCTGCATATCGCCAAAAAACTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
Stx2_phage_Min27_EU311208           AGGGGCAGTACGGAACCGTGGGGCAGGGGCTGCATATCGCCAAAAAACTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
Stx2_Sakai_BA000007                 AGGGGCAGTACGGAACCGTGGGGCAGGGGCTGCATATCGCCAAAAAACTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
Stx2_933W_AP004402                  AGGGGCAGTACGGAACCGTGGGGCAGGGGCTGCATATCGCCAAAAAACTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
STX2_C227-11                        AGGGGCAGTACGGAACCGTGGGGCAAGGGCTGCATATCGCCAAAAAATTGCTGCCGTTTATACCGGCGAATGCGGGCATTCTGCTGGTTCCGTGCTGTCG
                                    ************************* ********************* ****************************************************

STX2_PHI_272_PHAGE_HQ424691         TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
Stx2_converting_phage_AP005154      TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
Stx2_phage_Min27_EU311208           TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
Stx2_Sakai_BA000007                 TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
Stx2_933W_AP004402                  TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
STX2_C227-11                        TGGTGGTTCAGCGTTCACCACCGGAGCCGATGGCACATACAGTGACGCGAGTGGTGCCTCGGAGAATTCAACCCGCTGGGGTGTGGACAAGCCGCTGTAT
                                    ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691         AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAACCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGACGGCA
Stx2_converting_phage_AP005154      AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAATCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGGCGGTA
Stx2_phage_Min27_EU311208           AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAATCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGGCGGTA
Stx2_Sakai_BA000007                 AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAATCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGGCGGTA
Stx2_933W_AP004402                  AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAATCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGGCGGTA
STX2_C227-11                        AAGGACCTTATCGGTCGAACAAAAGCAGCACTGAAGAAGAATCCGAAAAATGTGCTGTTTGCCGTGGTGTGGATGCAGGGGGAATTTGATTTTGGCGGTA
                                    ***************************************** **************************************************** *** *

STX2_PHI_272_PHAGE_HQ424691         CGCCCGGAAATCACGCAGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGACGGATATGGCAGGTCAGTGCGTTGGTGGCTCTGCTGGCGG
Stx2_converting_phage_AP005154      CGCCGGTAAATCACGCCGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGGCGGATATGGCAGGCCAGTGCGTCGGTGGCTCTGCTGGCGG
Stx2_phage_Min27_EU311208           CGCCGGTAAATCACGCCGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGGCGGATATGGCAGGCCAGTGCGTCGGTGGCTCTGCTGGCGG
Stx2_Sakai_BA000007                 CGCCGGCAAATCACGCAGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGGCGGATATGGCAGGTCAGTGCGTCGGTGGCTCTGCTGACGG
Stx2_933W_AP004402                  CGCCGGTAAATCACGCCGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGGCGGATATGGCAGGCCAGTGCGTCGGTGGCTCTGCTGGCGG
STX2_C227-11                        CGCCGGTAAATCACGCAGCACAGTTTGGTGCGCTGGTTGATAAATTCCGTGCAGACCTGGCGGATATGGCAGGTCAGTGCGTCGGTGGCTCTGCTGGCGG
                                    **** * ********* ****************************************** ************* ******** ************* ***

STX2_PHI_272_PHAGE_HQ424691         TGTTCCCTGGATCTGTGGAGATACGACGTATTTCTGGAAGCAGAAGAACGAATCCTCGTACCAGACGGTGTACGGCAGCTACAAAAACAAAACGGAAAAG
Stx2_converting_phage_AP005154      TGTTCCCTGGATATGCGGGGACACGACGTATTTCTGGAAGCAGAAGAACGAATCCACGTACCAGACGGTGTATGGCAGCTATAAAAACAAAACGGAAAAG
Stx2_phage_Min27_EU311208           TGTTCCCTGGATATGCGGGGACACGACGTATTTCTGGAAGCAGAAGAACGAATCCACGTACCAGACGGTGTATGGCAGCTATAAAAACAAAACGGAAAAG
Stx2_Sakai_BA000007                 TGTTCCCTGGATATGCGGGGACACGACGTATTTCTGGAAGCAGAAGAACGAATCCTCGTACCAGACGGTGTACGGCAGCTACAAAAACAAAACGGAAAAG
Stx2_933W_AP004402                  TGTTCCCTGGATATGCGGGGACACGACGTATTTCTGGAAGCAGAAGAACGAATCCACGTACCAGACGGTGTATGGCAGCTATAAAAACAAAACGGAAAAG
STX2_C227-11                        TGTTCCCTGGATATGCGGGGACACGACGTATTTCTGGAAGCAGAAGAACGAATCCACGTACCAGACGGTGTACGGCAGCTACAAAAACAAAACGGAAAAG
                                    ************ ** ** ** ********************************* **************** ******** ******************

STX2_PHI_272_PHAGE_HQ424691         AATATCCATTTCGTACCGTTCATGACCGATGAGAACGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGGTATTACG
Stx2_converting_phage_AP005154      AATATCCATTTCGTACCGTTCATGACGGATGAGAACGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGATATTACG
Stx2_phage_Min27_EU311208           AATATCCATTTCGTACCGTTCATGACGGATGAGAACGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGATATTACG
Stx2_Sakai_BA000007                 AATATCCATTTCGTACCGTTCATGACGGATGAGAATGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGATATTACG
Stx2_933W_AP004402                  AATATCCATTTCGTACCGTTCATGACGGATGAGAACGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGATATTACG
STX2_C227-11                        AATATCCATTTCGTACCGTTCATGACGGATGAGAACGGGGTGAATGTGCCGACGAACAAACCGGAAGAAGACCCGGACATTCCGGGTATCGGATATTACG
                                    ************************** ******** ******************************************************** *******


STX2_PHI_272_PHAGE_HQ424691         GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGTTCATGGGCTCGCCGCGGGATTATTTCCGACCGTCTGGC
Stx2_converting_phage_AP005154      GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGTTCATGGGCTCGCCGCGGGATTATTTCCGACCGTCTGGC
Stx2_phage_Min27_EU311208           GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGTTCATGGGCTCGCCGCGGGATTATTTCCGACCGTCTGGC
Stx2_Sakai_BA000007                 GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGCGCCTGGGCACGCCGTGGGATTATTTCCGACCGTCTGGC
Stx2_933W_AP004402                  GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGTTCATGGGCTCGCCGCGGGATTATTTCCGACCGTCTGGC
STX2_C227-11                        GTTCGAAATGGCGTGACAGCTCAGCCACCTGGACGTCACAGGACAGGGCGAGCCATTTCAGTTCATGGGCTCGCCGCGGGATTATTTCCGACCGTCTGGC
                                    *************************************************************  * ***** ***** ***********************

STX2_PHI_272_PHAGE_HQ424691         AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
Stx2_converting_phage_AP005154      AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
Stx2_phage_Min27_EU311208           AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
Stx2_Sakai_BA000007                 AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
Stx2_933W_AP004402                  AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
STX2_C227-11                        AACGGCGATTTTGCGCCATGCGGGAAGAGTGGCGCTAAACGCGGGGGCATCATCGACAGTATCAGAGGTGCGCCCGTCATCGCCTTCCGGTGCAGAAGCC
                                    ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691         ACAGGCGTCACAACACTGCTCTCTTACCTTGTCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
Stx2_converting_phage_AP005154      ACAGGCGTCACAACACTGCTCTCTTACCTTGCCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
Stx2_phage_Min27_EU311208           ACAGGCGTCACAACACTGCTCTCTTACCTTGCCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
Stx2_Sakai_BA000007                 ACAGGCGTCACAACACTGCTCTCTTACCTTGCCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
Stx2_933W_AP004402                  ACAGGCGTCACAACACTGCTCTCTTACCTTGCCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
STX2_C227-11                        ACAGGCGTCACAACACTGCTCTCTTACCTTGCCAGCGAGTCAGAGGGAAGCCTGAAAGTACAGGGATGGTCAGCCAGTGGCGGCAGGGCAGAAGTGGTCA
                                    ******************************* ********************************************************************

STX2_PHI_272_PHAGE_HQ424691         GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
Stx2_converting_phage_AP005154      GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
Stx2_phage_Min27_EU311208           GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
Stx2_Sakai_BA000007                 GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
Stx2_933W_AP004402                  GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
STX2_C227-11                        GCGATGCGGAGGGAACCGGAGGTAAGGCAGTGAAGCTGACCAAGGAAGCCGGTAAAAGCAGCTGGGTGCTGGAGTACGCCGCGGGCAACGGTGCGGCTCT
                                    ****************************************************************************************************

STX2_PHI_272_PHAGE_HQ424691         GTTACAGAAAGGGGGGCAGATTCGCTGCCGCTTTAAGGTTTCGGGAGCGCTGGCTGCGAACCAGTATGTTATGGCGTTTTACTGGCCGGTATCTTCACTG
Stx2_converting_phage_AP005154      GTTACAGAAAGGGGGGCAGATTCGCTGCCGCTTTAAGGTTTCGGGAGCGCTGGCTGCGAACCAGTATGTTATGGCGTTTTACTGGCCGGTATCTTCACTG
Stx2_phage_Min27_EU311208           GTTACAGAAAGGGGGGCAGATTCGCTGCCGCTTTAAGGTTTCGGGAGCGCTGGCTGCGAACCAGTATGTTATGGCGTTTTACTGGCCGGTATCTTCACTG
```


Stx2_Sakai_BA000007
Stx2_933W_AP004402
STX2_C227-11
STX2_PHI_272_PHAGE_HQ424691
Stx2_converting_phage_AP005154
Stx2_phage_Min27_EU311208
Stx2_Sakai_BA000007
Stx2_933W_AP004402
STX2_C227-11
[1270 bp insertion]
Hypothetical protein

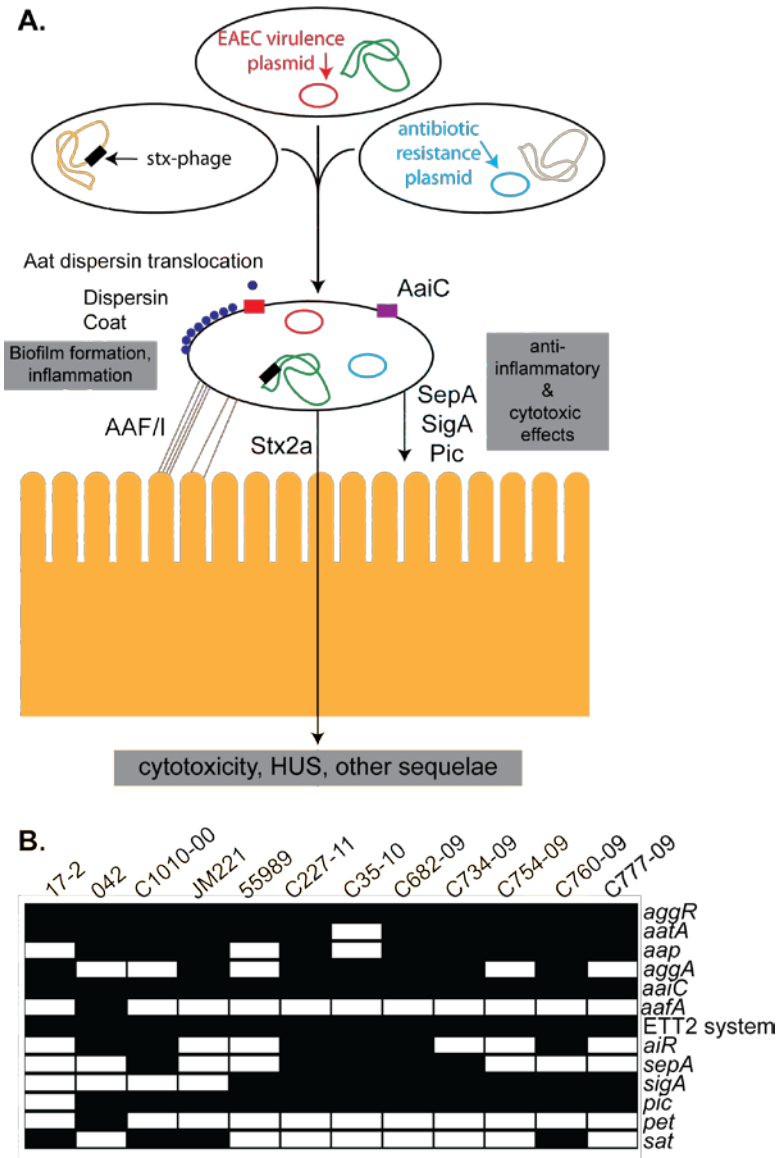


Supplementary Figure 8. Multiple sequence alignment of the *stx* gene region in lambdoid prophages found in multiple EHEC strains and in C227-11. The region includes the anti-termination protein Q (shaded grey), putative late promoter pR' (shaded light green), the Shiga-toxin 2 subunit A gene *stx2A* (shaded red), the Shiga-toxin 2 subunit B gene *stx2B* (shaded orange), and a hypothetical protein adjacent to the *stx2B* gene (shaded green). This region is highly conserved in the EHEC strains and in C227-11. In O157:H7 strains, exposure to antibiotics that trigger the bacterial SOS response results in increased production of

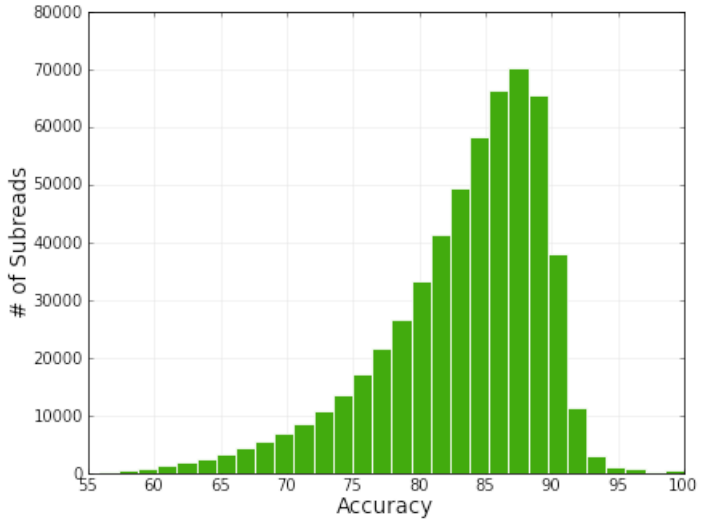
Shiga toxin. Experiments with several EHEC strains, which routinely contain highly homologous lambdoid phages to the Stx2-encoding phage in C227-11, have demonstrated that this increase in toxin production is due to activation of the Pr' promoter. The Pr' promoter upstream of *stx2* in C227-11 is identical to that in EDL933, suggesting that Stx2 production by C277-11 will likewise be inducible with antibiotics.



Supplementary Figure 9. Quantitative Reverse Transcriptase Polymerase Chain Reaction (qRT-PCR) to examine the induction of the Shiga-toxin genes by ciprofloxacin. The impact of antibiotics on C227-11 was examined using ciprofloxacin. Triplicate biological samples had RNA extracted after 6 hours of growth in LB media. The RNA was quantified and the expression level for the *stx2B* and *rpoA* gene were examined as previously described by described by Zhang et al [26] and Rasko et al [27] on biological duplicates. The graph demonstrates the increase in the expression level of the *stx2B* gene, using the *rpoA* gene for normalization. The level of expression in the strain not exposed to ciprofloxacin was normalized at 1 and the culture exposed to ciprofloxacin demonstrates an ~83 fold increase in gene expression. Error bars represent the standard deviation calculated using six datapoints (biological triplicates and technical duplicates for each).



Supplementary Figure 10. Proposed model for the evolutionary origins of and increased virulence associated with the O104:H4 isolates from the German outbreak. **A)** Notable steps in the evolution of the outbreak strain included acquisition of an Stx2-phage (black box) and a plasmid containing the CTX-M beta lactamase gene (blue), as well as additional virulence factors, by an ancestral EAEC isolate containing an EAEC virulence plasmid (red). We hypothesize that collectively these factors enable the EAEC to bind to and remain closely associated with the intestinal epithelium, which may promote increased uptake of Stx2 into the blood stream. **B)** EAEC strains encode diverse assemblages of virulence factors. EAEC isolate names are listed across the top of the figure; filled boxes mark genes that are present. Note that the outbreak strain, C227-11, encodes an unusual assortment of SPATE components, including SepA, SigA and Pic.



Supplementary Figure 11. Single pass read (subread) accuracy distribution for the continuous long read data generated for the C227-11 isolate. The mean accuracy of this distribution is 84.6%. However, given the long subread lengths generated in this study, the left tail of this distribution is longer than realized previously (Supplementary Methods), due to the fact that very long reads can be mapped with extremely high precision even at low accuracy. Therefore, one consequence of longer reads is reduced mean accuracy. As a result, a more accurate representation of the accuracy distribution is the mode, which in the C227-11 distribution shown here is at 88.2% accuracy.

Supplementary Table 1. PacBio *RS* continuous long-read sequencing statistics for the 12 *E. coli* isolates (from the isolates described in Table 1 of the main text). The number of post-filter reads (2nd column) represents the number of raw reads used in the analysis after filtering out low accuracy (< 75%) and short reads (< 100bp). The number of mapped reads (3rd column) indicate how many of the filtered reads from column 2 could be mapped to the TY-2482 or 55989 genomes. The mean (or mode) single-pass accuracy represent the mean (or mode) of the single pass accuracy distribution, while the consensus accuracy column represents the accuracy of the consensus sequence. The mean read length and 95th percentile read length represent the mean raw sequence read length and the read length of raw reads at the 95th percentile of the raw read length distribution (the read length distribution is exponential), respectively. Mean depth of coverage provides the amount of fold-coverage achieved for each genome. The per SMRTcell sequencing time provides the time in minutes it took to carry out a single sequencing run per SMRTcell, the number of SMRTcells indicates the number of SMRTcells run to get at the mean depth of coverage reported, and the final column is simple the product of the previous two columns, where minutes were then converted to hours.

Strain	# of post-filter reads	# of mapped reads	Single-pass accuracy Mean (%)‡	Consensus accuracy (%)	Mean readlength (bp)	95th percentile readlength (bp)	Mean SMRTBell readlength (bp)	95th percentile SMRTBell readlength (bp)	Mean depth of coverage	Per SMRTcell Seq Time (min)	# SMRT cells	Total Time for Collection (hours)
O42	495,931	215,414	84.2 (87.2)	99.998	1923	5003	2527	6829	75.3	90	9	15.75
17-2	254,237	148,330	*	NA	1953	5096	2503	6710	52.5	90	6	10.5
JM221	267,603	175,551	*	NA	2271	5795	2810	7351	73.4	90	8	14
C1010-00	407,475	294,567	*	NA	1873	4697	2376	6364	100.8	90	12	21
55989	186,794	162,912	84.4 (88.3)	99.999	2142	5260	2572	6713	63.4	90	6	10.5
C227-11	648,177	475,926	85.0 (88.2)	99.975	2191	5711	2900	7811	190.7	120	24	48
C35-10	362,535	223,253	*	NA	1818	4818	2196	6080	73.3	90	12	21
C682-09	257,792	195,513	*	NA	2003	5120	2287	6078	70.2	90	12	21
C734-09	328,472	154,014	*	NA	2135	5866	2369	6462	59.6	90	10	17.5
C754-09	166,002	80,917	*	NA	2169	5851	2364	6378	31.5	90	13	22.75
C760-09	537,878	254,893	*	NA	2094	5523	2411	6306	96.7	90	14	24.5
C777-09	140,689	74,932	*	NA	2255	5886	2714	7330	31.1	90	11	19.25
C227-11 (CCS)	419,589	416,656	97.8 (99.9)	NA	430	740	3076	8527	35.05	90	32	56

* denotes accuracies that could not be computed without a reference sequence. Similar chemistries used for all samples should result in little accuracy variation.
‡ The percentage indicated in parentheses, when present, represents the mode of the single-pass accuracy distribution
NA Indicates measurement not applicable

Supplementary Table 2. Single reads spanning > 95% of the 1,549bp plasmid at least five times. The read id corresponds to the id in the SRA submission of the C227-11 sequence data. The CCS length indicates the size of the insert in the SMRTbell construct for the indicated read. The number of passes indicates the number of times the DNA polymerase enzyme traversed the SMRTbell construct for the indicated read. The full sequence length of the read indicates how many bases were sequenced in the read, and the CCS accuracy indicates the percent of nucleotides in the pTY-3 plasmid that matched the consensus sequence of the consensus sequence for the indicated read (the accuracy is not expected to be 100% given real sequence differences that may exist in this plasmid between the TY-2482 and C227-11 genomes).

Read ID	CCS Length	# Passes	Readlength	CCS Accuracy (vs. TY-2482)
54845	1507	6	10333	97.4%
28873	1510	6	9266	98.1%
11743	1524	7	11744	98.4%
30326	1538	6	9654	99.5%
18647	1540	6	9688	99.6%
61841	1554	5	8913	99.2%

Supplementary Table 3. Comparison of publicly available genome assemblies from German outbreak strain isolates to TY-2482. Different genome assemblies from different outbreak isolates and related strains were compared against reference genomes (listed in column 1). The genomes being compared are given in column two (see Table 1 of main text for details). The two C227-11 assemblies listed are described in the Supplementary Methods. The Coverage column indicates the percentage of nucleotides in the reference assembly covered by the indicated strain, and the Identity column indicates the percentage of nucleotides that are identical between the indicated strain and the indicated reference strain. Under the assumption that the resequencing pipeline approximates the true sequence for C277-11, we can conservatively estimate the accuracy of our *de novo* assembly of C227-11 as 99.97, given that 0.02-0.03 of the identity difference to TY-2482 likely represents true variation between strains.

Data	Coverage	Identity	Alignment/Snp Calling
H1121	97.90%	99.98%	Nucmer; Show-snps
LB226	98.88%	99.90%	Nucmer; Show-snps
C227-11 de novo assembly	99.67%	99.95%	Nucmer; Show-snps
C227-11 resequencing	99.96%	99.97%	Nucmer; Show-snps
C227-11 resequencing	100.00%	99.98%	Blasr; EviCons

Supplmentary Table 4. Isolates in Whole Genome Phylogeny. Genomes for 53 E. coli strain isolates were used to produce Figure 2 in the main text. The id's, pathotype, and Genbank accession number for these 53 isolates are provided in this table.

Isolate	Pathotype	Genbank Accession
E. coli 536	ExPEC	CP000247.1
E. coli B7A	ETEC	AAJT00000000
E. coli ABU 83972	ExPEC	CP001671
E. coli APEC O1	ExPEC	CP000468.1
E. coli ATCC 8739	ExPEC	CP000946.1
E. coli BL21	Lab/Commensal	AM946981
E. coli CFT073	ExPEC	AE014075.1
E. coli ED1	Lab/Commensal	CU928162.2
E. coli E24377A	ETEC	CP000800.1
E. coli H10407	ETEC	FN649414
E. coli HS	Lab/Commensal	CP000802.1
E. coli IAI1	Fecal	CU928160.2
E. coli IAI39	ExPEC	CU928164.2
E. coli IHE3034	ExPEC	CP001969
E. coli O103:H2 str. 12009	EHEC	AP010958.1
E. coli O111:H- str. 11128	EHEC	AP010960.1
E. coli O127:H6 str. E2348/69	EPEC	FM180568.1
E. coli O157:H7 str. EC4115	EHEC	CP001164.1
E. coli O157:H7 str. EDL933	EHEC	AE005174.2
E. coli O157:H7 str. Sakai	EHEC	BA000007.2
E. coli O26:H11 str. 11368	EHEC	AP010953.1
E. coli O55:H7 str. CB9615	EHEC	CP001846.1
E. coli SE11	Fecal	AP009240.1
E. coli SMS-3-5	Environmental	CP000970.1
E. coli UMN026	ExPEC	CU928163.2
E. coli UTI89	ExPEC	CP000243.1
E. coli K12 DH10B	Lab/Commensal	CP000948.1
E. coli K12 W3110	Lab/Commensal	AP009048.1
E. coli B171	EPEC	AAJX00000000
E. coli 55989	EAEC	CU928145.2
E. coli 042	EAEC	FN554766
E. coli 101-1	EAEC	AAMK00000000
E. coli TY-2482	EAEC	TBD
E. coli LB226692	EAEC	AFOB00000000
E. coli H112180280	EAEC	TBD
E. coli 17-2	EAEC	TBD
E. coli JM221	EAEC	TBD
E. coli C1010-00	EAEC	TBD
E. coli C35-10	EAEC	TBD
E. coli C682-09	EAEC	TBD
E. coli C734-09	EAEC	TBD
E. coli C754-09	EAEC	TBD
E. coli C760-09	EAEC	TBD
E. coli C777-09	EAEC	TBD
E. coli C227-11	EAEC	TBD
E. coli 53638	EIEC	AAKB00000000
S. boydii CDC 3083-94	*Shigella*	CP001063.1
S. boydii Sb227	*Shigella*	CP000036.1
S. dysenteriae Sd197	*Shigella*	CP000034.1
S. dysenteriae 1012	*Shigella*	AAMJ00000000
S. flexneri 2a str. 2457T	*Shigella*	AE014073.1
S. flexneri 5 str. 8401	*Shigella*	CP000266.1
S. sonnei Ss046	*Shigella*	CP000038.1

Supplmentary Table 5. Known virulence factors mapped to the EAEC isolates sequenced in this study, with an indicator of presence in other EAEC isolates sequenced as described in the main text. The genome sequence (described in columns B-F) for the virulence factor (column A describes the factor and column B describes the original isolate from which the factor was identified) was mapped against all of the raw sequence reads for each of the twelve strains sequenced in this study. The number of sequences to which the factor sequence was aligned was then normalized by the factor's sequence length (the longer the sequence, the more hits we would expect) and used as a measure to determine presence of a given factor in a given genome. These adjusted counts are given in columns H-R for the twelve strains we sequenced. An adjusted count > 0 indicates the presence of a virulence factor in the associated strain. Mappings to individual raw reads (rather than consensus sequences) were carried out to enhance the sensitivity of identifying factors in the different genomes, given the long read lengths achieved via the SMRT sequencing runs.

UNIQID	Original isolate	Gene name	start_coo	end_coor	GenBank Accession Number	Inclusion in list	55989	17-2	O42	JM221	C1010-00	C35-10	C682-09	C734-09	C754-09	C760-09	C777-10	C227-11
aafA	Ec042	EC042_pAA048	39652	40134	FN554767	AggR regulated	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AafA_Ec042_FN554767.1	Ec042	EC042_pAA048	39652	40134	FN554767	AggR regulated	0.0	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aafB_Ec042_FN554767.1	Ec042	EC042_pAA030	22994	22554	FN554767	AggR regulated	0.0	0.0	9.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aafC	Ec042	EC042_pAA031	25533	23011	FN554767	AggR regulated	0.0	0.0	13.3	0.0	0.0	16.8	0.0	0.0	0.0	0.0	0.0	0.0
aafD_Ec042_FN554767.1	Ec042	EC042_pAA046	38499	39248	FN554767	AggR regulated	0.0	0.0	13.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aai_pathogenicity_island	Ec042	ECO42_4562 to ECO42_4577			FN554767	AggR regulated	31.5	33.9	15.3	32.2	17.9	4.5	12.2	14.0	30.2	8.1	25.3	12.5
aatA	Ec042_pAA	EC042_pAA008	5647	6885	FN554767	AggR regulated	9.9	7.1	6.2	3.2	3.2	3.0	0.0	12.0	14.6	5.9	8.6	5.8
aatB	Ec042_pAA	EC042_pAA009	6965	7603	FN554767	AggR regulated	0.8	0.0	0.0	0.0	0.3	0.0	0.0	1.0	0.0	0.6	0.0	0.0
aatC	Ec042_pAA	EC042_pAA010	7596	8225	FN554767	AggR regulated	1.6	1.5	1.1	2.0	0.3	0.8	0.0	2.4	1.4	0.3	0.8	0.2
aatD	Ec042_pAA	EC042_pAA011	8237	9451	FN554767	AggR regulated	7.9	9.0	5.1	6.5	3.0	3.0	0.0	11.0	13.7	6.7	11.3	4.8
agg3a_55989	55989	agg3a_55989	4258	4755	AF411067.1	AggR regulated	11.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.2	4.5	7.6	0.0
agg3B_55989	55989	agg3B_55989	3643	4083	AF411067.1	AggR regulated	2.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	1.0	3.6	0.0
agg3C_55989	55989	agg3C_55989	1083	3629	AF411067.1	AggR regulated	28.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	29.4	9.4	26.5	0.0
agg3D_55989	55989	agg3D_55989	277	1029	AF411067.1	AggR regulated	14.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.9	7.5	18.5	0.0
agg4a_C1010-00	C1010-00	agg4a_C1010-00	3780	4292	EU637023.1	AggR regulated	0.0	0.0	0.0	0.0	7.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
agg4B_C1010-00	C1010-00	agg4B_C1010-00	3278	3706	EU637023.1	AggR regulated	0.0	0.0	0.0	0.0	2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
agg4c_C1010-00	C1010-00	agg4c_C1010-00	733	3261	EU637023.1	AggR regulated	0.0	0.0	0.0	0.0	15.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
agg4D_C1010-00	C1010-00	agg4D_C1010-00	1	684	EU637023.1	AggR regulated	0.0	0.0	0.0	0.0	10.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
aggA		aggA				AggR regulated	0.0	0.0	0.0	0.0	0.0	0.4	0.0	1.1	0.0	0.0	0.0	0.5
aggA_17-2_U12894.1	17-2	aggA_17-2_U12894.1	3985	4500	U12894	AggR regulated	0.0	16.6	0.0	3.9	0.0	3.4	0.0	10.1	0.0	3.1	0.0	7.6
aggB_17-2_U12894.1	17-2	aggB_17-2_U12894.1	3435	3440	U12894	AggR regulated	0.0	1.8	0.0	4.8	0.0	2.3	0.0	11.5	0.0	4.1	0.0	6.4
aggC_17-2_U12894.1	17-2	aggC_17-2_U12894.1	905	3433	U12894	AggR regulated	0.0	22.7	0.0	16.4	0.0	8.8	0.0	29.4	0.0	10.6	0.0	14.4
aggD_17-2_U12894.1	17-2	aggD_17-2_U12894.1	133	891	U12894	AggR regulated	0.0	9.6	0.0	12.3	0.0	6.3	0.0	8.3	0.0	5.4	0.0	7.6
aggR	Ec042_pAA	EC042_pAA052	41877	41080	FN554767	AggR regulated	0.0	0.4	0.0	0.4	0.5	0.0	0.0	0.0	0.0	0.3	0.7	0.0
aggR gene	55989	aggR	48186	47389	CU928159.2	AggR regulated	21.7	19.4	9.1	13.4	0.0*	10.1	0.0	30.1	27.1	15.8	17.9	18.2
capU	Ec042	capU	1403	2224	AF134403	AggR regulated	3.9	3.9	11.2	0.0	0.0	5.7	2.2	3.4	3.0	4.0	3.3	22.3
gi|284919779:4251642-4251935 I	Ec042	EC042_4006	990336	990043	FN554766	AggR regulated	0.6	2.9	24.1	2.2	2.6	1.1	1.2	1.4	0.0	12.4	0.0	0.0
gi|284919779:c3401924-3401280	Ec042	EC042_3184	1840054	1840698	FN554766	AggR regulated	0.0	0.0	9.3	0.0	10.9	0.0	0.0	0.0	0.0	0.0	0.0	3.3
gi|284920982|emb|CBG34047.1|	Ec042	EC042_1227	3937251	3937652	FN554766	AggR regulated	29.5	33.1	26.7	45.5	30.6	30.1	19.6	23.4	28.1	26.9	31.9	0.0*
gi|284921996|emb|CBG35074.1|	Ec042	EC042_2242	2906559	2903440	FN554766	AggR regulated	13.2	48.0	49.2	46.1	54.6	22.1	25.8	28.4	19.9	36.1	20.6	1.3
gi|284921997|emb|CBG35075.1|	Ec042	EC042_2243	2903325	2900803	FN554766	AggR regulated	13.4	56.7	40.8	43.1	57.9	16.4	19.2	18.2	14.6	43.4	16.9	0.0
gi|284921999|emb|CBG35077.1|	Ec042	EC042_2244A	2900193	2899960	FN554766	AggR regulated	6.7	32.0	26.4	22.1	31.3	7.5	11.5	7.9	15.6	18.6	9.8	27.1
gi|284922918|emb|CBG36007.1|	Ec042	EC042_3181	1842773	1843042	FN554766	AggR regulated	5.0	4.6	20.6	7.6	6.8	3.5	12.0	9.5	10.4	1.7	4.6	44.2
gi|284922919|emb|CBG36008.1|	Ec042	EC042_3182	1842056	1842694	FN554766	AggR regulated	6.1	8.4	14.1	9.9	11.5	0.0	0.0	0.0	0.0	3.9	4.5	24.3
gi|284922920|emb|CBG36009.1|	Ec042	EC042_3183	1840839	1842071	FN554766	AggR regulated	9.6	14.3	24.1	20.7	14.9	2.5	8.0	12.7	12.2	6.3	11.3	12.3
gi|284922922|emb|CBG36012.1|	Ec042	EC042_3187	1836409	1838196	FN554766	AggR regulated	10.4	18.1	36.0	23.1	13.8	3.9	14.5	19.6	17.4	9.4	11.6	12.1
gi|284924231|emb|CBG37331.1|	Ec042	EC042_4509	404698	405153	FN554766	AggR regulated	6.7	12.3	13.7	28.7	2.8	3.6	15.5	16.9	6.4	14.0	15.0	0.0
gi|284924466|emb|CBG37594.1|	Ec042	EC042_4772	122940	123761	FN554766	AggR regulated	13.5	22.3	16.1	0.0	0.0	9.7	17.7	16.8	13.7	13.1	9.2	11.4
gi|284924582:17497-18447_VirK	Ec042_pAA	EC042_pAA023	17497	18447	FN554767	AggR regulated	10.0	14.6	26.4	0.0	0.0	12.7	7.5	10.3	11.9	11.9	12.7	18.1
gi|284924584|emb|CBG27756.1|	Ec042_pAA	EC042_pAA003	1397	2425	FN554767	AggR regulated	7.0	14.9	4.7	8.6	4.7	4.2	0.0	11.3	10.1	8.9	4.7	14.3
espP	EPEC	espP	55788	59690	HM138194	SPATE	0.0	0.0*	0.0*	0.0	0.0*	0.0	0.0	0.0*	0.0	0.0	0.0	14.5
pet	Ec042_pAA	EC042_pAA035	28073	31960	FN554767	SPATE	20.4	0.0*	0.0*	0.0	0.0*	4.3	17.7	23.4	24.5	11.4	18.2	8.4
pic	O42		4928958	4924840	FN554766	SPATE	39.4	0.0	26.4	28.6	27.2	9.9	54.1	59.6	41.0	26.8	37.3	8.5
sat	CFT073	c3619	3460261	3456362	AE014075	SPATE	0.0	29.0	0.0	0.0	0.0	27.2	0.0	0.0	0.0	0.0	0.0	4.8
sepA	S. flexneri M90T	sepA	262	4362	Z48219	SPATE	37.0	27.6	0.0	17.1	53.5	32.9	0.0	87.2	21.7	17.4	23.0	0.0
air	55989	air	3567798	3566278	CU928145		12.0	15.1	19.8	23.4	15.5	18.7	16.2	18.5	11.8	16.8	13.0	53.1
astA	55989	astA	53198	53082	CU928159.2		0.9	2.3	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
EAST1	EAEC	EAST1	61	177	L11241													0.0
gi|284924585|emb|CBG27757.1|	Ec042_pAA	EC042_pAA004	2429	2968	FN554767		7.9	7.6	4.1	2.2	4.5	4.5	0.0	7.8	8.1	3.7	7.3	16.1
rmoA	N/A	aida	1947	5810	GU810159		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	65.1
sisAB	CFT073	ShiA	3408019	3409062	AE014075		0.0	14.7	0.0	5.8	3.7	0.3	0.0	0.0	0.0	6.8	0.0	0.0
ydiE(hemP)	55989	hemP	1938245	1938436	CU928145		31.7	38.9	61.8	65.9	47.5	50.5	50.1	35.6	34.5	36.9	31.2	59.7

* indicates evidence for truncated or internally rearranged sequence

Supplementary Table 6. Investigation of the prevalence of unique regions from C227-11 when compared to 55989 and other EAEC isolates. Similar to the virulence factor mapping described in Supplementary Table 5, we mapped raw reads from the twelve EAEC strains we sequenced to the complete German outbreak reference genome indicated in column A (TY-2482) to identify regions that were not covered by sequence from the different EAEC strains. Those regions for which at least one of the non-C227-11 strains had an adjusted coverage count of 0 (as described in Supplementary Table 5 legend) are included in this table, with coordinates for the region with respect to the complete TY-2482 genome given in columns B-D.

location	end5	end3	span	55989	17-2	O42	JM221	C1010-00	C35-10	C682-09	C734-09	C754-09	C760-09	C777-10	C227-11
ECO_TY2482_	2503	3705	1202	0.0	0.0	0.0	41.5	0.0	32.4	23.6	50.3	0.0	0.0	0.0	31.5
ECO_TY2482_	11303	12343	1040	0.0	0.0	0.0	0.0	0.0	0.9	2.9	3.4	5.3	0.0	0.0	3.1
ECO_TY2482_	17084	20547	3463	0.0	0.0	0.0	0.0	0.0	2.9	12.7	12.9	20.9	0.0	0.0	10.4
ECO_TY2482_	27752	28688	936	0.0	0.0	0.0	0.0	0.0	6.0	19.0	24.0	17.3	8.1	0.0	23.9
ECO_TY2482_	53083	61507	8424	28.2	0.0*	0.0*	76.3	0.0*	45.2	74.6	85.4	47.8	40.6	29.9	105.3
ECO_TY2482_	65178	65360	182	0.0	0.0	0.0	0.0	0.0	5.7	39.2	33.0	13.9	3.7	0.0	36.1
ECO_TY2482_	286624	286810	186	0.0	0.0	0.0	0.0	0.0	28.8	8.4	4.4	23.8	19.3	4.1	7.3
ECO_TY2482_	287043	296211	9168	17.6	4.2	29.6	31.6	9.5	17.0	61.9	60.9	43.2	26.0	25.4	76.3
ECO_TY2482_	620641	622915	2274	0.0	0.0	0.0	0.0	0.0	4.6	17.0	21.2	6.8	0.0	0.0	19.4
ECO_TY2482_	623083	654701	31618	0.0	0.0	0.0	0.0	0.0	12.6	60.8	62.1	42.2	0.0*	0.0	130.3
ECO_TY2482_	659276	660180	904	0.0	0.0	0.0	0.0	0.0	2.2	15.5	18.0	14.2	0.0	0.0	43.3
ECO_TY2482_	661476	664083	2607	0.0	0.0	0.0	0.0	0.0	3.8	22.1	25.0	15.4	0.0	0.0	21.6
ECO_TY2482_	665999	666203	204	0.0	0.0	0.0	0.0	0.0	6.7	19.6	16.0	7.2	0.0	0.0	11.2
ECO_TY2482_	667251	673883	6632	0.0	0.0	0.0	0.0	0.0	7.2	24.7	26.1	17.0	0.0	0.0	20.9
ECO_TY2482_	674381	675692	1311	0.0	0.0	0.0	0.0	0.0	2.7	5.2	5.3	10.4	0.0	0.0	7.7
ECO_TY2482_	676270	679111	2841	11.5	11.8	0.0	0.0	11.2	19.4	30.5	31.8	23.4	7.9	10.9	32.5
ECO_TY2482_	679652	680554	902	0.0	0.0	0.0	0.0	0.0	3.8	11.0	14.0	13.6	0.0	0.0	20.3
ECO_TY2482_	812530	814798	2268	0.0	18.7	0.0	0.0	0.0	4.8	16.8	18.9	0.0	0.0	0.0	21.3
ECO_TY2482_	815833	816156	323	0.0	0.0	0.0	0.0	0.0	3.7	13.8	12.7	0.0	0.0	0.0	12.9
ECO_TY2482_	817534	818235	701	0.0	29.8	0.0	0.0	22.6	8.0	24.3	26.4	0.0	0.0	0.0	32.8
ECO_TY2482_	838266	838497	231	0.0	0.0	0.0	0.0	0.0	8.8	37.5	33.2	0.0	0.0	0.0	41.9
ECO_TY2482_	846777	849429	2652	0.0	0.0*	0.0	0.0*	0.0	17.1	26.0	74.3	0.0	0.0	0.0	59.0
ECO_TY2482_	1009501	1009697	196	0.0	0.0	0.0	0.0	0.0	8.9	34.8	30.1	15.9	12.7	0.0	51.0
ECO_TY2482_	1033035	1033520	485	0.0	0.0	0.0	63.6	6.8	2.2	8.5	9.3	8.5	0.0	0.0	12.8
ECO_TY2482_	1041020	1042984	1964	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.6
ECO_TY2482_	1043286	1044976	1690	0.0	11.9	0.0	0.0	0.0	9.6	10.8	11.1	18.7	0.0	0.0	11.0
ECO_TY2482_	1048219	1048558	339	0.0	0.0	0.0	0.0	0.0	4.2	27.6	16.3	13.0	0.0	0.0	20.8
ECO_TY2482_	1054379	1055342	963	0.0	0.0	0.0	0.0	0.0	2.5	11.4	5.8	8.8	0.0	0.0	11.2
ECO_TY2482_	1055829	1057962	2133	0.0	0.0	0.0	0.0	0.0	4.4	18.4	21.1	11.8	9.5	0.0	19.1
ECO_TY2482_	1607642	1612536	4894	0.0	0.0	0.0	0.0	0.0	8.4	16.7	20.1	24.9	0.0*	0.0	15.5
ECO_TY2482_	1612868	1613597	729	0.0	0.0	0.0	0.0	0.0	2.5	11.1	12.9	11.4	0.0	0.0	16.2
ECO_TY2482_	1615085	1615252	167	0.0	0.0	0.0	0.0	0.0	10.4	24.9	9.3	13.2	0.0	0.0	22.3
ECO_TY2482_	1623348	1623727	379	0.0	0.0	0.0	0.0	0.0	8.4	37.3	26.6	28.1	0.0	0.0	37.3
ECO_TY2482_	1624852	1625137	285	0.0	0.0	0.0	0.0	0.0	6.3	17.2	0.0	32.4	8.8	0.0	25.3
ECO_TY2482_	3122324	3122924	600	0.4	0.0	0.0	0.0	23.8	4.3	21.2	23.6	11.2	0.0	0.0	27.8
ECO_TY2482_	3126230	3132643	6413	0.0	0.0*	0.0	0.0	0.0*	54.7	30.8	32.6	26.2	0.0*	0.0*	38.4
ECO_TY2482_	3890422	3890553	131	0.0	0.0	0.0	0.0	0.0	1.7	14.5	9.5	10.4	0.0	0.0	11.6
ECO_TY2482_	5034313	5034466	153	0.8	0.0	3.0	0.0	0.0	1.6	2.5	1.9	0.0	0.0	1.5	1.3
ECO_TY2482_	5167116	5172982	5866	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	23.5
ECO_TY2482_	5173202	5180938	7736	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	20.8
ECO_TY2482_	5275737	5277251	1514	0.0	0.0	0.0	0.0	0.0	13.1	5.3	6.8	0.0	0.0	0.0	6.7
ECO_TY2482_	17318	18088	770	0.0	0.0	0.0	0.0	0.0	12.4	0.0	45.5	0.0	0.0	0.0	29.7
ECO_TY2482_	27598	28672	1074	0.0	0.0	0.0	16.8	0.0	18.7	7.0	3.3	10.9	23.0	14.9	18.4
ECO_TY2482_	30824	32331	1507	0.0	0.0	0.0	0.0	27.9	13.6	0.0	49.9	0.0	0.0	0.0	24.1
ECO_TY2482_	37400	37532	132	0.0	0.0	0.0	0.0	0.0	3.3	0.0	0.0	0.0	0.0	0.0	3.9
ECO_TY2482_	37730	38107	377	0.0	0.0	0.0	0.0	0.0	82.0	0.5	45.8	0.0	0.0	0.0	55.3
ECO_TY2482_	44056	88695	44639	0.0	0.0	0.0	0.0	0.0	147.0	0.0*	141.1	0.0	0.0	0.0	71.1
ECO_TY2482_	25440	30951	5511	0.0*	0.0*	0.0	0.0*	37.4	19.4	0.0*	57.5	0.0*	0.0*	0.0*	37.9
ECO_TY2482_	45503	46711	1208	0.0	0.0	0.0	0.0	0.0	9.1	3.0	40.1	0.0	0.0	0.0	31.4

ECO_TY2482_	46954	47215	261	0.0	34.1	0.0	0.0	0.0	21.4	0.0	18.6	0.0	31.7	0.0	12.7
ECO_TY2482_	54025	55002	977	0.0	20.1	0.0	0.0	0.0	18.4	0.0	44.6	0.0	0.0	0.0	28.3
ECO_TY2482_	57890	59742	1852	0.0	19.5	0.0	21.8	0.0	12.5	0.0	32.1	0.0	0.0	0.0	19.9
ECO_TY2482_	60511	61335	824	0.0	9.2	0.0	9.1	0.0	5.3	0.0	15.5	0.0	8.2	0.0	5.8
ECO_TY2482_	70380	70609	229	0.0	0.0	0.0	0.0	0.0	8.2	0.0	14.3	0.0	0.0	0.0	8.4
ECO_TY2482_	71067	71626	559	0.0	0.0	0.0	0.0	0.0	9.3	0.0	31.0	0.0	0.0	0.0	18.0
ECO_TY2482_	0	1549	1549	0.0	0.6	0.0	0.0	0.0	3.2	36.3	8.6	0.0	0.0	0.0	25.9

* indicates evidence for truncated or internally rearranged sequence

Supplementary Table 7. Putative virulence factors present in C227-11. A condensed version of Supplementary Table 5 highlighting the virulence factors identified in C227-11. Nucleotide sequences corresponding to these factors are provided in the 'Accession' column (GenBank accession numbers).

Gene	Location	Putative virulence function	Described	Accession
Master regulator				
aggR	pAA plasmid	Master regulator of a package of EAEC plasmid virulence genes, including aggregative adherence factors, fimbriae AAF/I-AAF/IV, and a large cluster of genes inserted on a pathogenicity island at the PheU locus [1]	EAEC-042	Z18751
AggR regulated				
aatPABCD	pAA plasmid	Encodes ABC protein responsible for transporting the dispersin protein out of the outer membrane of EAEC [2]	EAEC-042	AY351860
aap	pAA plasmid	Encodes a 10 kDa secreted protein named dispersin, and is responsible for 'dispersing' EAEC across the intestinal mucosa	EAEC-042	Z32523
aggABCD	pAA plasmid	Encodes AAF/I mediates adherence to colonic mucosa and hemagglutination of erythrocytes [4]	EAEC-JM221	AY344586
aaiA-P	Chromosome	Encodes a type VI secretion system encoded on the *pheU* island on the chromosome. Mode of action unknown [5]	EAEC-042	-
Toxin gene				
stx2a	Chromosome	Shiga toxin (Stx); A-B-type toxin that inhibits protein synthesis in eukaryotic cells, is thought to be required for the manifestations of EHEC infection, such as hemorrhagic colitis	*E. coli* O157:H7 EDL933	X07865
sigA	Chromosome	Encodes an IgA protease-like homologue [8]	*S. flexneri* 2a 2457T	NC_004337
pic	Chromosome	Encodes the Pic protein; mucinase activity and is capable of causing hemagglutination of erythrocytes [9]	*S. flexneri* 2a & EAEC-042	AF097644
sepA	pAA plasmid	Encodes *Shigella* extracellular protein. May induce mucosal atrophy and tissue inflammation in *S. flexneri* [10]	*S. flexneri* 2a & EAEC-C1010-00	Z48219
Other genes				
air	Chromosome	Possible aggregation and adherence [11]	EAEC-042	-
capU	pAA plasmid	Hexosyltransferase homologue [6]	EAEC-042	AF134403
ETT2 genes	Chromosome	Putative Type III secretion system	EAEC-O42	-

1. Bernier C, Gounon P, Le Bouguenec C. Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative Escherichia coli as a sensitive probe for detecting the AAF-encoding operon family. Infect Immun 2002;70:4302-11.
2. Nishi J, Sheikh J, Mizuguchi K, et al. The export of coat protein from enteroaggregative Escherichia coli by a specific ATP-binding cassette transporter system. J Biol Chem 2003;278:45680-9.
3. Sheikh J, Czeczulin JR, Harrington S, et al. A novel dispersin protein in enteroaggregative Escherichia coli. J Clin Invest 2002;110:1329-37.

4. Nataro JP, Deng Y, Maneval DR, German AL, Martin WC, Levine MM. Aggregative adherence fimbriae I of enteroaggregative Escherichia coli mediate adherence to HEp-2 cells and hemagglutination of human erythrocytes. Infect Immun 1992;60:2297-304.
5. Dudley EG, Thomson NR, Parkhill J, Morin NP, Nataro JP. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative Escherichia coli. Mol Microbiol 2006;61:1267-82.
6. Czeczulin JR, Whittam TS, Henderson IR, Navarro-Garcia F, Nataro JP. Phylogenetic analysis of enteroaggregative and diffusely adherent Escherichia coli. Infect Immun 1999;67:2692-9.
7. Karmali MA. Infection by verocytotoxin-producing Escherichia coli. Clin Microbiol Rev 1989;2:15-38.
8. Rajakumar K, Sasakawa C, Adler B. Use of a novel approach, termed island probing, identifies the Shigella flexneri she pathogenicity island which encodes a homolog of the immunoglobulin A protease-like family of proteins. Infect Immun 1997;65:4606-14.
9. Henderson IR, Czeczulin J, Eslava C, Noriega F, Nataro JP. Characterization of pic, a secreted protease of Shigella flexneri and enteroaggregative Escherichia coli. Infect Immun 1999;67:5587-96.
10. Benjelloun-Touimi Z, Si Tahar M, Montecucco C, Sansonetti PJ, Parsot C. SepA, the 110 kDa protein secreted by Shigella flexneri: two-domain structure and proteolytic activity. Microbiology 1998;144 (Pt 7):1815-22.
11. Sheikh J, Dudley EG, Sui B, Tamboura B, Suleman A, Nataro JP. EilA, a HilA-like regulator in enteroaggregative Escherichia coli. Mol Microbiol 2006;61:338-50.

Supplementary Table 8. EAEC-specific virulence factors mapped to the O104 and other EAEC reference isolates sequenced in our study. Similar to Supplementary Table 5 but focused on EAEC-specific virulence factors present in the AA plasmid from the O42 strain.

UNIQID	55989.0	17-2	O42	JM221	C1010-00	C35-10	C682-09	C734-09	C754-09	C760-09	C777-10	C227-11
EC042_pAA001 noGene -- transposase (pseudogene)	1.6	183.4	61.3	1.9	41.4	60.3	0.0	3.7	0.0	36.6	1.4	1.7
EC042_pAA005 noGene -- hypothetical protein	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA005A noGene -- conserved hypothetical protein	0.0	1.0	0.0	1.0	2.4	3.1	0.0	6.8	0.0	0.6	0.0	4.5
EC042_pAA005A noGene -- transposase (pseudogene)	4.0	1.4	1.0	2.1	8.7	2.1	1.7	2.7	5.7	3.5	5.2	2.8
EC042_pAA007 aatP -- permease	7.9	7.1	2.5	2.6	2.5	3.7	0.0	5.4	6.6	4.4	6.7	4.1
EC042_pAA011 noGene -- transposase (pseudogene)	16.7	31.6	24.7	9.2	26.7	48.9	24.6	42.9	4.8	31.3	18.2	46.4
EC042_pAA013 noGene -- transposase	1.1	18.1	5.6	0.0	4.0	6.7	6.1	14.2	0.9	5.0	1.0	11.9
EC042_pAA014 noGene -- transposase	15.3	31.9	30.5	0.0	16.4	23.0	14.0	38.4	0.0	12.0	9.4	30.4
EC042_pAA015 noGene -- hypothetical protein	0.0	0.0	3.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA016 noGene -- site-specific recombinase	19.8	12.4	19.9	17.6	26.3	48.3	0.0	25.0	36.1	33.3	25.8	19.7
EC042_pAA017 ccdB -- post-segregation toxin (cytotoxic protein)	11.1	10.8	1.6	8.6	12.8	22.5	0.0	15.8	10.7	16.2	7.5	10.1
EC042_pAA018 ccdA -- toxin addiction system antidote protein	2.8	4.2	2.5	0.8	6.6	18.8	0.0	0.8	6.9	3.8	0.0	3.0
EC042_pAA018 noGene -- conserved hypothetical protein (pseudogene)	0.0	0.0	4.1	0.0	0.0	8.3	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA020 noGene -- conserved hypothetical protein	0.0	0.0	9.7	0.0	0.0	58.8	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA022 noGene -- glycosyl transferase	10.5	13.3	22.8	0.0	0.0	10.8	7.9	14.6	13.5	15.1	12.0	8.2
EC042_pAA023 noGene -- conserved hypothetical protein (pseudogene)	113.2	22.7	57.6	41.5	81.3	25.7	48.3	85.6	144.4	130.6	112.1	91.7
EC042_pAA023 noGene -- transposase (pseudogene)	19.2	2.4	11.8	1.6	15.9	6.6	8.4	13.6	27.8	22.7	17.7	13.5
EC042_pAA023 virK -- virulence protein required for expression/correct membrane localisation o	8.3	9.8	18.3	0.0	0.0	11.5	5.4	5.1	8.9	8.8	7.9	5.9
EC042_pAA026 noGene -- transposase	0.0	0.0	12.4	0.0	0.0	20.3	0.0	0.0	0.0	27.2	0.0	0.0
EC042_pAA027 noGene -- transposase	0.0	0.8	0.6	0.0	1.0	1.0	1.0	0.8	0.0	0.2	0.0	0.4
EC042_pAA028 noGene -- transposase	0.0	2.1	0.0	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.3
EC042_pAA028 noGene -- transposase (pseudogene)	0.0	0.0	5.4	0.0	0.0	0.0	0.0	0.0	0.0	9.9	0.0	0.0
EC042_pAA030 aafB -- afimbrial adhesin	0.0	0.0	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA031 aafC -- aggregative adherence fimbria II usher protein	2.0	0.0*	20.6	2.0	1.8	17.9	0.0	1.9	1.0	1.6	1.5	1.4
EC042_pAA031 afaB -- fimbrial chaperone protein (pseudogene)	0.0	0.0	4.8	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA033 noGene -- transposase	127.8	28.7	75.3	107.6	44.1	192.2	104.2	93.4	82.2	89.8	83.8	72.3
EC042_pAA034 noGene -- transposase	479.2	141.2	342.0	631.4	188.1	785.2	364.5	410.8	357.4	325.5	322.3	383.6
EC042_pAA035 noGene -- transposase (pseudogene)	1.0	159.9	62.2	17.0	52.8	44.5	21.2	0.0*	1.0	6.2	0.0*	15.9
EC042_pAA035 pet -- serine protease (plasmid-encoded toxin Pet)	0.0*	0.0*	19.2	0.0*	0.0*	0.0	0.0*	0.0*	0.0	0.0*	0.0*	0.0*
EC042_pAA039 noGene -- transposase	0.0	0.0	2.5	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA041 noGene -- conserved hypothetical protein	0.0	8.6	1.1	10.7	0.0	6.0	0.0	17.8	0.0	0.0	0.0	8.5
EC042_pAA042 noGene -- conserved hypothetical protein	0.0	26.0	6.1	23.5	0.0	12.9	0.0	42.0	0.0	0.0	0.0	28.1
EC042_pAA042 noGene -- transposase (pseudogene)	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA046 aafD -- chaperone protein	1.5	0.5	8.9	0.9	0.0	0.0	0.0	0.0	0.6	0.0	0.7	0.2
EC042_pAA047 noGene -- hypothetical protein	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA048 aafA -- major fimbrial subunit of aggregative adherence fimbria II	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA048 noGene -- transposase (pseudogene)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA051 noGene -- transposase	7.4	11.3	23.8	30.1	5.9	32.3	16.8	18.1	13.4	14.5	5.7	10.7
EC042_pAA052 aggR -- transcriptional activator	27.5	20.9	6.3	19.5	10.3	10.9	0.0	32.8	31.4	16.6	18.7	23.3
EC042_pAA052 noGene -- transposase (pseudogene)	1.2	2.9	1.6	0.7	0.0	0.6	0.0	1.4	1.0	0.8	2.2	0.6
EC042_pAA056 noGene -- hypothetical protein	3.0	0.9	0.0	0.9	0.0	4.2	0.0	6.2	0.0	1.6	0.0	4.9
EC042_pAA056 noGene -- transposase (pseudogene)	0.0	0.0	7.3	0.0	3.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA059 noGene -- transposase	0.0	12.1	6.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA060 noGene -- conserved hypothetical protein	7.1	8.8	0.7	6.1	0.0	15.4	0.0	23.6	7.2	4.0	4.0	17.4
EC042_pAA061 noGene -- conserved hypothetical protein	20.8	34.6	14.2	46.1	20.7	13.3	9.5	45.1	29.2	19.6	19.8	48.1
EC042_pAA062 noGene -- transposase	0.3	17.1	20.6	72.2	14.7	10.9	0.0	47.9	0.4	7.2	0.0	37.6
EC042_pAA063 noGene -- transposase	0.0	1.3	1.0	5.4	1.2	1.0	0.0	3.2	0.0	1.5	0.0	2.1
EC042_pAA064 noGene -- transposase	0.0	8.7	16.6	45.6	12.4	10.7	0.0	29.1	0.0	6.6	0.0	29.8
EC042_pAA064 noGene -- transposase (pseudogene)	0.0	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA066 finO -- fertility inhibition protein (conjugal transfer repressor)	28.4	12.2	6.8	16.4	12.0	70.4	0.0	13.7	36.0	28.5	48.5	11.9
EC042_pAA067 traX -- conjugative transfer protein	0.0	0.0	11.7	0.0	15.9	35.8	0.0	19.8	0.0	21.8	16.5	19.9
EC042_pAA068 traI -- DNA helicase I	0.0	0.0	36.7	0.0	40.3	126.5	0.0*	55.5	0.0	50.6	37.3	42.6
EC042_pAA069 traD -- conjugative transfer protein	0.0	0.0	23.5	0.0	22.4	68.6	0.0	0.0	0.0	27.2	26.4	0.0
EC042_pAA070 noGene -- putative conjugative transfer protein	0.0	0.0	6.2	0.0	7.2	25.4	0.0	0.0	0.0	5.0	5.5	0.0
EC042_pAA071 traT -- enterobacterial complement resistance protein	0.0	0.0	11.1	0.0	11.8	77.7	0.0	0.0	0.0	13.5	7.4	0.0
EC042_pAA072 traS -- conjugative transfer protein	0.0	0.0	3.0	0.0	0.0	4.3	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA073 traG -- conjugative transfer protein	0.0	0.0	27.4	0.0	31.7	84.5	0.0	0.0	0.0	36.4	29.0	0.0
EC042_pAA074 traH -- conjugative transfer protein	0.0	0.0	11.7	0.0	18.0	53.8	0.0	0.0	0.0	19.7	21.3	0.0
EC042_pAA075 trbf -- conjugative transfer protein	0.0	0.0	2.9	0.0	0.0	1.9	0.0	0.0	0.0	1.1	4.8	0.0
EC042_pAA076 trbJ -- conjugative transfer protein	0.0	0.0	0.6	0.0	2.6	5.1	0.0	0.0	0.0	0.4	0.0	0.0
EC042_pAA077 trbB -- conjugative transfer protein	0.0	0.0	13.3	0.0	19.5	48.8	0.0	0.0	0.0	12.5	11.3	0.0
EC042_pAA078 traQ -- conjugative transfer protein	0.0	0.0	3.3	0.0	2.7	15.3	0.0	0.0	0.0	3.4	2.2	0.0
EC042_pAA079 trbA -- conjugative transfer protein	0.0	0.0	3.6	0.0	5.3	4.7	0.0	0.0	0.0	6.9	10.2	0.0

EC042_pAA080 traF -- conjugative transfer protein	0.0	0.0	11.0	0.0	14.8	53.1	0.0	0.0	0.0	23.2	29.0	0.0
EC042_pAA081 trbE -- conjugative transfer protein	0.0	0.0	1.2	0.0	2.3	2.4	0.0	0.0	0.0	0.4	0.0	0.0
EC042_pAA082 traN -- conjugative transfer protein	0.0	0.0	22.5	0.0	28.9	74.7	0.0	0.0	0.0	20.0	6.6	0.0
EC042_pAA083 trbC -- conjugative transfer protein	0.0	0.0	7.1	0.0	8.6	25.8	0.0	0.0	0.0	9.3	15.6	0.0
EC042_pAA084 noGene -- putative conjugative transfer protein	0.0	0.0	0.0	0.0	0.0	8.1	0.0	0.0	0.0	2.8	2.1	0.0
EC042_pAA085 traU -- conjugative transfer protein	0.0	0.0	17.0	0.0	26.1	52.0	0.0	0.0	0.0	27.3	34.6	0.0
EC042_pAA086 traW -- conjugative transfer protein	0.0	0.0	6.4	0.0	9.2	21.3	0.0	0.0	0.0	9.0	5.2	0.0
EC042_pAA087 trbI -- conjugative transfer protein	0.0	0.0	3.4	0.0	2.3	11.2	0.0	0.0	0.0	4.0	2.9	0.0
EC042_pAA088 traC -- conjugative transfer protein	0.0	0.0	23.1	0.0	34.3	86.0	0.0	0.0	0.0	41.0	34.1	0.0
EC042_pAA089 traR -- conjugative transfer protein	0.0	0.0	1.3	0.0	6.5	20.6	0.0	0.0	0.0	5.7	7.6	0.0
EC042_pAA090 traV -- conjugative transfer protein	0.0	0.0	8.3	0.0	30.9	49.0	0.0	0.0	0.0	17.5	19.1	0.0
EC042_pAA091 trbD -- conjugative transfer protein	0.0	0.0	2.1	0.0	0.0	7.6	0.0	0.0	0.0	0.4	0.0	0.0
EC042_pAA092 traP -- conjugative transfer protein	0.0	0.0	8.7	0.0	4.1	10.9	0.0	0.0	0.0	5.4	4.0	0.0
EC042_pAA093 traB -- conjugative transfer protein	0.0	0.0	18.0	0.0	21.4	60.6	0.0	0.0	0.0	25.2	23.2	0.0
EC042_pAA094 traK -- conjugative transfer protein	28.7	0.0	10.5	23.2	11.6	41.2	0.4	0.0	48.7	38.4	42.5	0.0
EC042_pAA095 traE -- conjugative transfer protein	10.3	0.0	5.1	8.9	14.1	27.0	0.0	0.0	16.5	15.2	22.9	0.0
EC042_pAA096 traL -- conjugative transfer protein	6.4	0.0	2.7	0.7	3.8	16.8	0.0	0.0	6.8	5.6	6.4	0.0
EC042_pAA097 traA -- pilin precursor	6.4	0.0	4.9	5.2	5.5	18.5	0.0	0.0	6.3	10.8	5.9	0.0
EC042_pAA098 traY -- conjugative transfer protein	2.0	0.0	0.0	2.5	5.0	0.0	0.0	0.0	2.3	3.3	6.2	0.0
EC042_pAA099 traJ -- conjugative transfer protein	1.2	0.0	14.7	0.0	0.0	0.0	0.0	0.0	0.7	11.1	25.1	0.0
EC042_pAA100 traM -- conjugative transfer protein	12.5	0.0	2.4	3.8	3.8	8.8	0.0	12.9	13.9	9.8	15.4	6.8
EC042_pAA101 noGene -- putative transglycosylase	8.6	0.0	3.3	7.2	11.3	37.4	0.0	17.6	12.1	16.0	17.5	7.4
EC042_pAA102 noGene -- conserved hypothetical protein	27.1	0.0	20.7	17.0	19.0	82.6	0.0	31.9	29.8	41.4	40.0	20.5
EC042_pAA103 noGene -- conserved hypothetical protein	0.0	0.0	3.8	0.0	0.0	22.8	0.0	0.0	0.0	3.3	6.6	0.0
EC042_pAA103 noGene -- conserved hypothetical protein (pseudogene)	8.7	8.5	8.0	7.1	0.0	13.6	0.0	0.0	11.6	18.1	14.9	0.0
EC042_pAA103 noGene -- putative transposase (pseudogene)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA106 hok -- post-segregation killing protein	0.0	0.0	0.0	0.0	1.2	0.8	0.0	0.0	0.0	0.6	0.0	0.0
EC042_pAA107 mok -- modulator of Hok protein	0.0	0.0	1.3	0.0	5.1	5.9	0.0	0.0	0.0	1.5	2.6	0.0
EC042_pAA108 noGene -- conserved hypothetical protein	0.0	0.0	2.7	0.0	0.0	9.2	0.0	0.0	0.0	2.8	9.5	0.0
EC042_pAA109 psiA -- plasmid SOS inhibition protein	16.0	0.0	15.4	0.0	20.3	48.6	0.0	0.0	13.4	29.6	28.8	0.0
EC042_pAA110 psiB -- PsiB protein	0.0	0.0	3.2	0.0	3.6	12.9	0.0	1.7	0.0	6.1	7.2	2.7
EC042_pAA111 noGene -- ParB-like nuclease	0.0	0.0	26.1	16.6	25.6	99.5	0.0	35.3	0.0	31.6	22.9	27.6
EC042_pAA112 noGene -- conserved hypothetical protein	0.0	0.0	0.6	0.8	2.4	7.5	0.0	1.6	0.0	0.9	1.2	0.7
EC042_pAA113 noGene -- conserved hypothetical protein (pseudogene)	0.0	1.8	2.6	8.7	9.9	8.7	0.0	0.0	0.0	5.9	2.6	0.0
EC042_pAA113 ssb -- single-stranded DNA-binding protein	0.0	0.0	11.9	9.8	9.1	33.4	0.0	11.5	0.0	10.8	4.9	2.2
EC042_pAA116 noGene -- conserved hypothetical protein	9.1	6.4	10.1	7.1	17.5	12.8	0.0	0.0	9.7	20.3	21.5	0.0
EC042_pAA120 noGene -- conserved hypothetical protein	0.0	0.0	9.1	0.0	13.2	12.5	0.0	0.0	0.0	10.8	11.1	0.0
EC042_pAA121 noGene -- conserved hypothetical protein	0.0	18.3	21.2	0.0	25.1	30.1	0.0	0.0	0.0	25.4	31.1	0.0
EC042_pAA122 noGene -- conserved hypothetical protein	0.0	0.0	6.2	0.0	3.0	5.1	0.0	0.0	0.0	8.9	0.0	0.0
EC042_pAA123 noGene -- hypothetical protein	0.0	0.0	0.8	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA124 noGene -- conserved hypothetical protein	0.0	2.1	0.0	3.2	0.0	0.8	0.0	0.0	0.0	1.2	0.0	0.0
EC042_pAA126 noGene -- conserved hypothetical protein	0.0	0.0	0.0	0.0	0.5	0.7	0.0	0.0	0.0	9.8	5.5	0.0
EC042_pAA127 noGene -- conserved hypothetical protein	0.0	0.0	9.8	0.0	7.3	3.7	0.0	0.0	0.0	7.4	11.3	0.0
EC042_pAA128 noGene -- conserved hypothetical protein	0.0	0.0	9.6	0.0	14.2	41.5	0.0	29.8	0.0	6.2	8.2	20.7
EC042_pAA129 noGene -- conserved hypothetical protein (pseudogene)	0.0	0.0	7.8	0.0	14.5	23.5	0.0	8.3	0.0	10.5	14.8	1.0
EC042_pAA129 noGene -- putative antirestriction protein	0.0	0.0	7.3	0.0	8.7	25.1	0.5	12.2	0.0	7.6	15.5	7.2
EC042_pAA131 noGene -- conserved hypothetical protein	0.0	0.0	28.1	0.0	25.6	68.0	0.0	38.6	0.0	16.9	24.4	28.9
EC042_pAA132 noGene -- conserved hypothetical protein	0.0	0.0	4.5	0.0	7.2	14.7	0.0	12.7	0.0	2.0	0.9	6.4
EC042_pAA133 noGene -- conserved hypothetical protein	0.0	0.0	0.6	0.0	2.0	3.9	0.0	2.4	0.0	0.5	0.0	0.0
EC042_pAA134 noGene -- conserved hypothetical protein (pseudogene)	0.0	0.0	1.2	0.0	17.2	74.5	0.6	21.7	0.0	7.5	8.2	13.1
EC042_pAA134 noGene -- DNA methylase	0.0	0.0	14.0	0.0	24.0	66.4	0.0	34.0	0.0	16.8	18.7	19.9
EC042_pAA136 parM -- plasmid segregation protein	0.0	0.0	20.9	0.0	0.0	12.1	0.0	0.0	0.0	21.1	27.9	0.0
EC042_pAA137 noGene -- transposase (pseudogene)	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA137 pinQ -- putative DNA invertase (pseudogene)	0.0	0.0	4.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA137 stbB -- plasmid partition protein	0.0	0.0	2.0	0.0	0.0	2.6	0.0	0.0	0.0	6.8	7.0	0.0
EC042_pAA141 noGene -- hypothetical protein	0.0	0.0	0.7	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA142 eitA -- iron ABC transporter. substrate-binding protein	0.0	0.0	17.0	0.0	0.0	9.1	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA142 eitB -- putative iron ABC transporter.permease protein (pseudogene)	0.0	0.0	1.1	0.0	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA142 eitC -- iron ABC transporter. ATP-binding protein (pseudogene)	0.0	0.0	5.0	0.0	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA145 eitD -- major facilitator superfamily protein	0.0	0.0	18.8	0.0	0.0	9.6	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA145 imm -- colicin-e2 immunity protein (microcin-e2 immunity protein)(pseudogene)	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA147 noGene -- putative plasmid stabilisation system protein	0.0	0.0	11.4	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA148 noGene -- putative plasmid stabilisation system protein	0.0	0.0	2.9	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA149 noGene -- conserved hypothetical protein	0.6	0.0	16.9	0.0	0.5	2.4	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA149 repA4 -- plasmid replication (pseudogene)	56.9	10.3	7.2	42.8	18.0	35.2	0.0	19.8	50.9	42.3	63.2	17.5
EC042_pAA153 noGene -- transposase (pseudogene)	0.0	0.0	11.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

EC042_pAA153 repA2 -- replication regulatory protein 1	0.0	0.0	4.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA155 noGene -- transposase (pseudogene)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
EC042_pAA157 noGene -- transposase	0.0	3.8	1.6	0.4	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.2
EC042_pAA158 noGene -- transposase	4.5	0.7	0.5	0.0	1.7	2.7	2.3	0.7	1.9	1.2	4.2	1.2
EC042_pAA158 noGene -- transposase (pseudogene)	0.0	0.0	3.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

* indicates evidence for truncated or internally rearranged sequence