Supporting Information

Eppinger et al. 10.1073/pnas.1107176108

SI Materials and Methods

Bacterial Strains. The strain panel represents a diverse and fairly comprehensive strain collection consisting of historical strains and strains associated with recent sporadic *E. coli* O157:H7 outbreaks obtained from clinical, bovine, and environmental sources. Strain history and associated metadata of the *E. coli* O157:H7 strains used in this study are listed in Dataset S1. The strain panel contains 208 lineage I/II strains linked to the spinach (SP) outbreak obtained from 24 of the 26 affected states in August and September, 2006. Strain sources are the bovine host reservoir, bagged spinach, and ill patients; these strains enabled the study of fine genetic polymorphisms from a potential animal reservoir to passage in human (Dataset S1). The bovine strains were obtained from cattle in the Salinas Valley of California in an area of spinach production implicated in the outbreak. Three strains, lineage I/II strains EC536 and EC508 and lineage II strain EC869, were included based on cladistic multilocus and sequence typing and microarray analysis that showed phylogenetic relatedness to typical spinach-outbreak strains (1). The panel contains four clinical strains from two outbreaks [Taco Bell (TB) and Taco John (TJ)] that occurred almost simultaneously in November and December, 2006, and further lineage I strains EDL933, Sakai, and TW14588, the lineage I/II strain TW14359, the lineage II strains FRIK2000 and FRIK966 (2), and 191 strains derived from the 2006 SP outbreak (Dataset S1). The original glycerol stocks have been cultivated only once for single-colony DNA isolation subjected to whole-genome sequencing. Bacterial strains are available from the Food and Drug Administration (FDA) *E. coli* reference strain collection.

Genome Architecture and Phage Prevalence. Genomic architectures were compared on the nucleotide and proteome level using Mauve (3, 4) and BLAST score ratio analysis (Table S2). For each of the predicted proteins in the O157:H7 genomes, a BLASTP raw score was obtained for the alignment against itself (REF_SCORE) and the most similar protein (QUE_SCORE) in each of the genomes. Dividing the QUE_SCORE obtained for each query genome protein by the REF_SCORE normalized these scores. This methodology also was applied for the comprehensive phage analysis. Proteins with a normalized ratio <0.4 were considered to be nonhomologous. A normalized BLAST score ratio of 0.4 generally indicates that two proteins are 30% identical over their entire length (5).

Genome Visualization. The χ^2s and GC skews were computed according to Nelson et al. (6). For the chromosomal χ^2 a window size of 2 kb and a sliding window of 1 kb were used, whereas a window size of 1 kb and a sliding window of 0.2 kb were used for the plasmids. GC skews were calculated with a window size of 1 kb for the chromosome and 0.2 kb for the plasmids. The whole-genome alignment tool NUCmer (7) was used to calculate the overall genome identities.

Sanger DNA Sequencing. Genomic DNA of 12 *E. coli* O157:H7 strains was subject to random shotgun sequencing and closure strategies using a combination of Sanger and 454 sequencing as previously described [EC4206 (7 contigs), EC4045 (8 contigs), EC4042 (4 contigs), EC4196 (186 contigs), EC4113 (231 contigs), EC4076 (135 contigs), EC4401 (186), EC4486 (165 contigs), EC4501 (250 contigs), EC508 (272 contigs), and EC869 (147 contigs)] (Dataset S1). (6). Two random insert pHOS2 libraries with insert sizes of 3–5 kb and 10–12 kb and a fosmid library of 35–40 kb were constructed. Draft genome sequences were assembled using Celera Assembler (8).

Genome Annotation. The chromosomes, plasmids, and draft contigs were annotated manually using the MANATEE system (http://manatee.sourceforge.net/).

454 DNA Sequencing. Whole-genome draft 454 sequences of six additional representative outbreak-associated strains [EC4009 (845 contigs), EC4084 (881 contigs), EC4127 (791 contigs), EC4191 (267 contigs), EC4205 (1583 contigs), EC4192 (958 contigs)] and the FDA reference strain EC536 (678 contigs) were obtained in collaboration with the National Bioforensics Analysis Center. Assembly of sequencing reads into contigs and subsequent ordering of these contigs into scaffolds was performed with GS De Novo Assembler software (Roche). For genome sequencing, two to five Roche FLX 454 pyrosequencing runs were typical, and each yielded, on average, 5 million bases.

Pyrosequencing Assays. Pyrosequencing assays were designed to confirm predicted SNP locations and to detect genotypic variation within the studied outbreak population. When a BLASTN search of the sequence surrounding the SNP (600 nt) resulted in multiple and ambiguous hits in some genomes, a pyrosequencing assay was not attempted. Verified SNP assays were used to test 208 strains from the SP outbreak and related strains for variation. Primer design was accomplished using PSQ Assay Design Software v 1.0.6 (Biotage AB), which designs forward and reverse PCR primers for a 50- to 150-bp amplicon including one biotinylated primer and a sequencing primer.

PCR for Pyrosequencing. Each PCR reaction contained 5 μL DNA template, 5 μL 10× PCR buffer with 1.5 mM $MgCl_2$ (Perkin-Elmer), 5 μL of 2.5 mM dNTP (Pharmacia), 1.3 μL forward primer (10 μM), 1.3 μL biotinylated reverse primer (10 μM), and 1.5 U Taq DNA polymerase (Promega) in a 50-μL reaction. Amplification was performed in a PTC-200 thermal cycler (MJ Research) under the following conditions: initial denaturation at 94 °C for 5 min; 94 °C for 30 s, 54 °C for 30 s, and 72 °C for 30 s (45 cycles); and final incubation at 72 °C for 10 min.

Pyrosequencing SNP Assay. A total of 20 μL of biotinylated PCR product was immobilized onto 3-μL streptavidin-coated Sepharose beads (Amersham Biosciences) in 40 μL of binding buffer, pH 7.6 (10 mM Tris·HCl, 2 M NaCl, 1 mM EDTA, 0.1% Tween 20), in each well of a 96-well plate. Each plate was incubated at room temperature for 10 min with shaking (900 rpm) to keep the beads dispersed. Beads were picked up with a vacuum prep tool (Biotage AB), immersed for 5 s each in 70% ethanol, 0.2 M NaOH, and washing buffer, pH 7.6 (1 mM Tris-acetate), and then were dispensed into a 96-well plate containing 4 μL 10 mM primer in 40 μL annealing buffer, pH 7.6 (20 mM Tris, 2 mM magnesium acetate-tetrahydrate). Annealing was performed by heating each plate to 80 °C in a heat block for 2 min. The pyrosequencing was performed on an automated PSQ96MA instrument using the PSQ 96 SNP reagent kit (Qiagen) according to the manufacturer's instructions.

Optical Mapping. Optical maps for nine strains were generated that facilitated closure and assembly, thus allowing a detailed study of the prophage dynamics and their respective genome localization in the studied population. Optical maps were prepared by OpGen. After gentle lysis and dilution, high-molecular-mass

genomic DNA molecules were spread and immobilized onto derivatized glass slides and digested with BamHI. Using the Argus instrument (Opgen), the DNA digests were stained with YOYO-1 fluorescent dye and were photographed using a fluorescent microscope interfaced with a digital camera. Automated image-analysis software (OpGen) located and sized fragments and assembled fragments from multiple scans into whole-chromosome optical maps.

Biolog Phenotype Microarray. Strains to be tested were plated on Biolog Universal Growth medium and incubated overnight at 37 °C. Cells were swabbed from the plates after overnight growth and were suspended in appropriate medium containing Dye Mix C (Biolog). Then 100 μL of a 1:200 dilution of an 85% transmittance suspension of cells was added to each well of the phenotype microarray (PM) plates. Plates 1–8, which test for catabolic pathways for carbon, nitrogen, phosphorus, and sulfur and for biosynthetic pathways, and plates 9 and 10, which test for osmotic/ion and pH effects, were used in this study. IF-0 GN base (Biolog) was used for PM plates 1 and 2. IF-0 GN base plus 20 mM sodium succinate, pH 7.1, and 2 μM ferric citrate was used for plates 3–8. IF-10 base (Biolog) was used for plates 9 and 10. Plates were incubated in the OmniLog (Biolog) for 48 h with readings taken every 15 min. Data analysis was performed using Kinetic and Parametric software (Biolog). Phenotypes were determined based on the difference of the area under the kinetic curve of dye formation between the mutant and wild type. Data points for the entire 48 h were used for PM plates 1–8, and area differences were mean-centered by plate.

Accession Numbers. The genome sequences are deposited in GenBank under accession nos. EC4115 [CP001164 (chromosome), CP001163 (pO157), CP001165 (pEC4115)], EC4206 (ABHK00000000, 54.362 reads, 7 contigs), EC4045 (ABHL00000000, 56,363 reads, 8 contigs), EC4042 (ABHM00000000, 54,604 reads, 4 contigs), EC4196 (ABHO01000000, 55,767 reads, 186 contigs, 9.11×), EC4113 (ABHP00000000, 54,692 reads, 231 contigs, 8.37×), EC4076 (ABHQ01000000, 58,353 reads, 135 contigs, 8.75×), EC4401 (ABHR01000000, 53,229 reads, 186 contigs, 8.09×), EC4486 (ABHS00000000, 57.454 reads, 165 contigs, 8.6×), EC4501 (ABHT00000000, 56.737 reads, 250 contigs, 9.18×), EC508 (ABHU00000000, 55,923 reads, 147 contigs), and EC869 (ABHU00000000, 54,466 reads, 147 contigs, 8.66×). The 454 sequence draft of strains EC4205 (ADVB01000000), EC4084 (ADUY01000000), EC4127 (ADUZ01000000, EC4191 (ADVA01000000), EC4192 (ADUX01000000), EC4192 (ADMX01000000), and EC536 (ADVC01000000) has been deposited at GenBank. Respective genome assemblies have been deposited in the National Centers for Biotechnology Information (NCBI) Assembly Archive, and the electropherogram data of the Sanger sequencing traces are available from the NCBI Trace Archive.

SNP Discovery. We have developed and applied a bioinformatics pipeline that facilitates the discovery of fine polymorphisms in completed and draft genomes to reconstruct a detailed evolutionary history of the *E. coli* O157:H7 lineage. SNP calls were curated based on quality scores and were curated manually for false-positive SNP calls resulting from misalignment, paralogous genes, or insufficient genome coverage. A total of 1,225 candidate SNPs in both coding and noncoding DNA were identified in pairwise genome comparisons in respect to the reference *E. coli* O157:H7 strain EC4115. SNPs were identified by comparing the closed chromosome of EC4115 with draft contigs of the outbreak-associated strains, reference strains, and the previously sequenced strains EDL933 (9), Sakai (10), TW14359 (11), and FRIK2000 and FRIK966 (2) using MUMmer (12). Because we had access to detailed genome assembly, for genomes sequenced in this study we considered the polymorphic site to be of high quality when its underlying sequence in the sequenced reference genome EC4115 and query genomes comprised at least three sequencing reads with an average Phred quality score >30 (13). For all other strains (sequenced by others), the deposited consensus sequences were used. Because of the highly repetitive nature of the *E. coli* O157:H7 genomes, the SNP data set was curated manually for false positives, and positions within repeats and laterally transferred regions were excluded from further phylogenetic analysis. Repetitive regions, such as phages and mobile genetic elements, and their respective integration loci, were identified using a combination of entries into the Phage Sequence databank (http://phage.sdsu.edu/~rob/cgi-bin/phage.cgi databank) and the RepeatMasker (http://www.repeatmasker.org) and PhageFinder (14) software packages. When a comprehensive all-versus-all BLASTN (15) search among the reference and query genomes of the sequence (40 nt) resulted in multiple, ambiguous, and paralogous hits in some genomes, these regions were excluded from further phylogenetic analyses. Alignment quality was validated further by generating ClustalX multisequence alignments using 600 nt surrounding the SNP. By mapping the position of the SNP to the annotation in the reference *E. coli* strain EC4115 genome, it was possible to determine the effect on the deduced polypeptide and to classify each SNP as intragenic (synonymous, nonsynonymous) or intergenic.

Multilocus Variable Number Tandem Repeat Analysis. Seven assays were used based on the Multiplex PCR assays described in ref. 16. Primers were obtained from Integrated Data Technologies, Inc. The primer sets were divided into three separate multiplex PCR reactions: TR1, TR5, and TR6; TR3, TR4, and TR7; and TR2. The PCR consisted of 3 μL 10× Buffer (Qiagen), 2 μL $MgCl_2$ (Qiagen), 2 μL of 2.5 mM dNTPs (GE Life Sciences) and 1 μL each primer, 0.5 μL HotStart Taq (Qiagen) in a 30-μL reaction with 1 μL of DNA template. Amplification was performed in a PTC-200 thermal cycler (MJ Research) under the following conditions: initial denaturation at 94 °C for 4 min; 94 °C for 45 s, 53–58 °C for 45 s, and 72 °C for 1 min (32 cycles); and final incubation at 72 °C for 5 min. Visualization of the products was conducted using the Agilent Bioanalyzer 2100 and agarose gel. The multilocus variable-number tandem repeat analysis codes for the seven tested loci are presented in Dataset S1.

Phylogenetic Analysis. The primary analysis was performed using the complete data of intra- and intergenic SNPs among the studied O157:H7 strains. The concatenated SNP data were analyzed by the HKY93 method (17) with 500 bootstrap replicates, and the results were used to generate a phylogenetic tree according to the PhyLM algorithms (18) using the Geneious software package (http://www.geneious.com) and SplitsTree4 (19). The genomes of seven additional strains associated with the SP outbreak [EC4084, EC4127, EC4191, EC4205, EC4192, and EC4009 (sequenced for this study)] and strains TW14359, as well as the genomes of two bovine strains FRIK2000 and FRIK966 (2), and the FDA reference strain EC536 were genotyped using the identified *E. coli* O157:H7 lineage SNP backbone comprising 1,225 positions.

SI Results and Discussion

Genotypic SNP Profiling. This panel contains a collection of 208 SP-outbreak strains comprising 191 clinical strains and 17 SP strains collected in August and September, 2006. Nine bovine strains from California farms and six TB strains (Dataset S3). collected within the same timeframe rounded out the test panel. Although the majority of SP strains (197/208) carry the typical outbreak (EC4205) SNP pattern of group A strains, 12 strains derived from various states showed different SNP patterns (Dataset S3). With one exception, within a state, when we observed an atypical SP strain at least one other strain exhibited the typical pattern. We note the important exception: In Maine the three clinical strains (EC4114, EC4115, and EC4116) showed an identical but

atypical SNP pattern compared with the dominant SP genotype. SNP-based genotyping supported the clustering of group A bovine strain from California into two distinct subtypes, in accordance with the established genotypes of the bovine type-strains EC4206 (4/11) and EC4196 (7/11) (Fig. 2). In three instances we noted that tested strains acquired mutations under laboratory-cultivation conditions when the DNA preparation used for sequencing was compared with the original DNA of the single-colony isolate. This finding is of major interest for forensic outbreak investigations. These fine genetic polymorphisms comprise two point mutations and a small inversion of 1.9 kbp.

Insertion Sequence Elements as Drivers of Prophage Microevolution. A copy of insertion sequence (IS) element IS*629* (ECH74115_2932, ECH74115_2933) disrupts key replication genes in the Shiga-like toxin c (*stx2c*)-converting prophages of the lineage I/II strains derived from the SP and TB outbreaks (Fig. 3), whereas *in silico* analysis showed that IS*629* is absent in the phylogenetically more distant lineage I/II strain EC508 and in *stxc2* prophages from other sources (Fig. 3). IS*629* (EC74115_3515, ECH74115_3516) also disrupts genes (EC4115_3414, ECH74115_3417) within the *sbcB*, occupying Stx2-prophage of lineage I/II strains, including EC508 (*sbcB*, ECH7EC508_0841), whereas in Stx2-positive lineage I strains this phage is found integrated at the *wrbA* locus and does not carry IS*629* (Fig. 1 *C* and *D* and Fig. S2*A*). The IS*629* insertion found within the potentially *stx1*-converting phage at the *yehV* locus (ECH74115_3231, ECH74115_3232) (Fig. S2*B*) is absent in the reference lineage I/II strain EC508 and in more distantly related lineage I outbreak strains and bovine lineage II strain EC869. More importantly, IS element profiles are in agreement with the SNP-derived phylogenetic placement of the strains and provide valuable genomic signatures for investigating the state of phage evolution in the *E. coli* O157:H7 lineage (Fig. 2 and Table S2).

Mutation in the *N*-Acetyl-D-Galactosamine/D-Galactosamine Phosphotransferase Transport System. SNP discovery revealed a single nonsynonymous nucleotide polymorphism in the *agaF* gene in strain EC4115 at position 4,124,954 (G:C to A:T) that distinguishes all analyzed SP and TB outbreak strains (Dataset S2). AgaF is part of an *N*-acetyl-D-galactosamine/D-galactosamine (Aga/Gam) phosphotransferase transport system (PTS), and thus these Aga-negative strains cannot use Aga as a primary carbon or nitrogen source (20). The base substitution in *agaF*, which encodes $EIIA^{Aga/Gam}$, changes a conserved glycine residue to serine (Gly91Ser), which has been demonstrated previously to underlie the Aga-negative phenotype in these strains (20).

Polymorphism in the Stx1 Prophage. Comprehensive analyses reveal an overall syntenic organization in key structural and enzymatic phage components, with the notable exception of several polymorphic regions (shaded in red in Fig. S2*B*) that most likely are the result of recombinatorial events caused by the homologous nature of the lambdoid phages. These regions code structural phage components (EDL933: Z3323–Z3331; TW14359: ECSP_2949/ECSP_2950; EC4115: ECH74115_3134); a methyltransferase; hypothetical proteins with no assigned function (Z3345–Z3353, ECSP_2964–ECSP_2971, ECH74115_3220–3227); an antiterminator, superinfection exclusion protein, and an ssDNA-binding protein (Z3361–Z3363); and a superoxide dismutase (Z3312) unique to EDL933. Comprehensive analyses revealed three polymorphic regions and identified IS*629* element insertions as microevolutionary drivers. Both SP outbreak-associated phage types carry fragmented *repP* phage replication pseudogenes, resulting from IS*629* element disruption, which may impair replication and immobilize these phages in the strains associated with SP outbreak. The smaller variant shows a disrupted phage portal protein (TW14359, ECSP_2949/ECSP_2950) compared with strain EC4115 (ECH74115_3134). The latter carries an additional joint phage-related 45,632-bp fragment at the tRNA Met locus integrated within the borders of the SP15-PPV phage (*intV*, ECSP_2997, ECH74115_3251) mediated by the integrase *intO* (ECH74115_3178). This insert appears to have arisen from a duplication event unique to Maine (ME) strain EC4115. The region is a complex composite and is organized syntenically to two joined loci of the SP9 prophage. Comprehensive analyses of corresponding phage loci in strains TW14359 and EC4115 reveal a single polymorphic region among the SP9 prophages (marked in red, Fig. S2*B*), tail fiber protein ECSP_1769 in TW14359 with numerous homologous genes in the lambdoid phages in strain EC4115 (ECH74115_2758, ECH74115_3188, ECH74115_2165, ECH74115_1203). ECH74115_1881 is a fragmented pseudogene with no assigned function. The adjacent tRNA loci may have facilitated double homologous recombination and rearrangements of these joined prophage fragments.

Non-Shiga Toxin–Carrying Phage Markers and Genomic Islands in *E. coli* O157:H7. ***P2-type prophage at the* yegH *gene locus.*** A P2-type prophage (31,943 bp) is integrated at the *yegH* gene locus in strain EC4115 and is flanked by 18-bp perfect direct repeats (Table S2). This phage is syntenic to *Enterobacteria* phage PP2 (33,593 bp, NC_001895) (Fig. S3*B*), and its integrase is homologous to P2 integrases, such as those of enterotoxigenic *E. coli* (ETEC) O148:H28 strain B7A (ZP_03029820) and avian pathogenic *E. coli* (APEC) O1:K1:H7 (21). Two polymorphic phage regions were identified in structural head and tail components as well as in a cluster of hypothetical proteins with no assigned function.

SP13 prophage length polymorphism. Three lineage I/II strains, the SP outbreak strain EC4076 and TB strains EC4401 and EC4486, show a unique deletion of the Sp13 prophage (21,118 bp) (Fig. S3*C*). The triplet tRNA Leu-Cys-Gly serves as a phage-integration site in Sp13-carrying strains, and phage integration is marked by a 121-bp perfect direct repeat; in strain EC4076 this sequence has been reduced to a single 121-bp sequence (Table S2).

Prophage remnant in lineage I/II strain EC4115. The lineage I/II strain EC4115 carries a prophage remnant (9,247 bp) that is integrated into the tRNA dihydrouridine synthase A gene (*dusA*) (ECH74115_5551) resulting in 104-bp imperfect direct repeats left and right (IDRL/R). This prophage codes for 21 genes (ECH74115_5529–ECH74115_5550) (Fig. S3*D* and Table S2). The integrase (ECH74115_5550) is highly similar to that of *E. coli* strain 101–1, an atypical enteroaggregative *E. coli* (EAEC) strain that was responsible for a large outbreak in Japan in the late 1990s (EC1011_3392). This strain contains few classic EAEC virulence factors, as determined by DNA hybridization; however, the nature of this strain suggests increased virulence over other EAEC strains. This prophage integration not only provides novel genetic material in strain EC4115 but also drives the microevolution of DusA, when comparing strains EDL933 and EC4115. The repeat sequence shares 84% identity with the EDL933 *dusA* (Z5647). The 104-bp IDRL sequence of the N-terminal duplicated *dusA* pseudogene (ECH74115_5529) is 100% identical to the *dusA* regions in EDL933 (ECH74115_5551); however the IDRR sequences (now the N-terminal region of the *dusA* gene in strain E4115) shares only 84% nucleotide identity. The observed phage-introduced polymorphisms result in DusA protein identities of 97%.

Prophage remnant in lineage I/II strain EC508. A P4-like prophage (14,596 bp) is integrated at the *dinD* locus (ECH7EC508_1640) of the human isolated lineage I/II strain EC508, resulting in 10-bp perfect direct repeats (Table S2). This phage codes for 24 genes (ECH7EC508_1615–ECH7EC508_1636) (Fig. S3*E*). *In silico* analysis shows that this phage scar also is present in the lineage I-derived TJ strains TW14588 (14,834 bp) and EC4501 (14,832 bp) and shows a length polymorphism of 3,769 bp.

Genomic Islands and Islets. Lineage I/II strains, except for strains EC508 and EC536 (Fig. S3*F*), feature a 15,147-bp deletion relative to lineage I strains. This genomic region neighboring the murein transglycosylase E (EC74115_1680) is flanked by 22-bp imperfect direct repeats, leaving a hybrid of the duplicated 22-bp sequence, ATTCTtgCgccgctgAgCGCCC, at the deletion site (Table S2). The insertion is mediated by a copy of IS element IS*629* (ECH74115_1957/1958) and introduces 13 genes. Noteworthy are a PTS-dependent system *treAdhkMLKR*, an iron(III) ABC transporter, and five additional hypothetical proteins of unknown function. This island may prove beneficial, countering iron limitations within the mammalian host and potentially increasing bacterial fitness. We noted also a length polymorphism in the lineage I/II island carrying strain EC508 compared with lineage I strains. EC508 carries an additional 3,155-bp fragment, which is an integral part of the adversely orientated regulator DhrR (EC1706 and ECH7EC508_3647), and the putative outer membrane autotransporter EngD (EC1707 and ECH7EC508_3648); the most likely secondary loss results in truncated protein products in lineage I strains (DhrR, 961 vs. 547 aa; EngD, 363 vs. 158 aa). Lineage II strains EC869, FRIK2000, and FRIK966 contain yet another genomic variation (ECH7EC869_0175, ECH7EC869_0178) that is likely derived from the lineage I island but appears to be in the process of decay and codes parts of the PTS-dependent dihydroxyacetone kinase operon DhrKH (ECH7EC869_0174, ECH7EC869_0175). This truncated 4,318-bp islet features two imperfect direct repeats, (ATTCTTGCCATCTAAGAATATCGCC, ATTCTACCGCCGCTGAGCGCCC). Lineage I/II strains contain a small 1,667-bp insertion (Fig. S3*G*) that also is present in strains EC508 and EC536. This region codes a cluster of Rhs-core proteins disrupted by a transposase. These ankyrin-repeat sequences are known virulence factors and have been implicated in mediating phage susceptibility (23, 24). *In silico* comparison of completed and draft genomes indicates that the tested lineage I and lineage II strains lack these lineage I/II-specific signatures (Fig. S3*G*).

Evidence for *E. coli* O157:H7 Microevolution Under Laboratory Conditions. In three instances we noted that tested strains acquired point mutations under laboratory cultivation conditions. That is, when the DNAs of single-colony strains chosen for sequencing were analyzed by pyrosequencing, the SNPs at positions 3,691,927 and 4,479,378 in strains EC4042 and EC4076, respectively, were readily verified (Dataset S2). However, when DNAs were prepared from original freezer stocks of these strains and tested with pyrosequencing, the SNPs could not be confirmed. For all other verified SNPs, DNA from the original glycerol stock was tested along with the single-colony isolated DNA, and no other polymorphisms were found. The third example of culture-induced polymorphism affects the genomic structure and is discussed below.

Inversion Affects Structural *Enterobacteria* Phage Components. *In silico* comparison of the completed lineage I/II genomes EC4115 and TW14359 revealed yet another polymorphic state mediated by 15-bp inverted direct repeats (IDR), *attR*/*L*. We found an inverted 1,9-kb region within the borders of *Enterobacteria* phage P2 harbored in these strains (Fig. 1 *C* and *D* and Fig. S4). Similar mechanistic structures mediate phase variation in enteric bacteria. In this particular case, however, the inversion results in two chimerical fusion proteins (ECH74115_3032 and ECH74115_3035) in strains EC4115 and TW14359 (Fig. S4). The observed alteration of the encoded surface-exposed structural proteins may interfere with host–pathogen interaction and immunity and thus impact pathogenic potential. Of note, in strains EDL933 and Sakai these signature IDRs are found in another genomic context within prophage H (Fig. S4). Genomic structures were validated with inversion-specific primer pairs (Table S1). Interestingly, the original glycerol stock of EC4115 showed no alteration in genomic structure in this region compared with SP outbreak strains EC4042 and the two other ME strains, EC4114 and EC4116. However, when the DNA preparation of the EC4115 single-colony isolate that served as template for genomic sequencing was used, the inversion could be confirmed readily by PCR (see Table S1 for primer sequences). Our findings provide examples of the types of fine genetic polymorphisms that can occur during routine laboratory cultivation and single-colony isolation.

Altered Sucrose Phenotype in *E. coli* O157:H7. Previous work showed that, unlike *E. coli* K12 strains, the majority of *E. coli* O157:H7 strains analyzed from the FDA culture collection mentioned above displayed a D-serine–negative/D-sucrose–positive phenotype (22) resulting from a sucrose utilization regulon integrated into the D-serine operon, *dsdCXA*. Targeted pyrosequencing analyses confirmed that all three SP outbreak strains collected from Maine (EC4114, EC4115, and EC4116) harbor the same point mutation, resulting in a truncated pseudogene product of 373 aa (from 415 aa), and support the Biolog-observed ME strain-specific D-sucrose–negative phenotype amino acids. We detected another nonsynonymous SNP unique to the three ME strains that affects a putative sucrose-specific PTS system, the *N*-acetylmuramic acid–specific EIIBC component (ECH74115_3659) at position 457. We note here that the SNP panel contains numerous polymorphisms that potentially alter physiological and metabolic capabilities in *E. coli* O157:H7. The SNP backbone described herein for the *E. coli* O157:H7 lineage provides the basis for further experimental validation.

1. Cebula TA, et al. (2005) Molecular applications for identifying microbial pathogens in the post-9/11 era. *Expert Rev Mol Diagn* 5:431–445.
2. Dowd SE, et al. (2010) Microarray analysis and draft genomes of two Escherichia coli O157:H7 lineage II cattle isolates FRIK966 and FRIK2000 investigating lack of Shiga toxin expression. *Foodborne Pathog Dis* 7:763–773.
3. Rissman AI, et al. (2009) Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* 25:2071–2073.
4. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 14:1394–1403.
5. Rasko DA, Myers GS, Ravel J (2005) Visualization of comparative genomic analyses by BLAST score ratio. *BMC Bioinformatics* 6:2.
6. Nelson KE, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323–329.
7. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
8. Huson DH, et al. (2001) Design of a compartmentalized shotgun assembler for the human genome. *Bioinformatics* 17(Suppl 1):S132–S139.
9. Perna NT, et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–533.
10. Hayashi T, et al. (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8:11–22.
11. Kulasekara BR, et al. (2009) Analysis of the genome of the *Escherichia coli* O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence. *Infect Immun* 77:3713–3721.
12. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30:2478–2483.
13. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175–185.
14. Fouts DE (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 34: 5839–5851.
15. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA* 87:2264–2268.
16. Noller AC, McEllistrem MC, Harrison LH (2004) Genotyping primers for fully automated multilocus variable-number tandem repeat analysis of *Escherichia coli* O157:H7. *J Clin Microbiol* 42:3908.
17. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22:160–174.
18. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online–a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* 33(Web Server issue):W557–559.

19. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.
20. Mukherjee A, Mammel MK, LeClerc JE, Cebula TA (2008) Altered utilization of *N*-acetyl-d-galactosamine by *Escherichia coli* O157:H7 from the 2006 spinach outbreak. *J Bacteriol* 190:1710–1717.
21. Johnson TJ, et al. (2007) The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol* 189:3228–3236.
22. Mukherjee A, Jackson SA, Leclerc JE, Cebula TA (2006) Exploring genotypic and phenotypic diversity of microbes using microarray approaches. *Toxicol Mech Methods* 16:121–128.
23. Liu K, Knabel SJ, Dudley EG (2009) *rhs* genes are potential markers for multilocus sequence typing of *Escherichia coli* O157:H7 strains. *Appl Environ Microbiol* 75: 5853–5862.
24. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R (2007) Genes under positive selection in *Escherichia coli*. *Genome Res* 17:1336–1343.

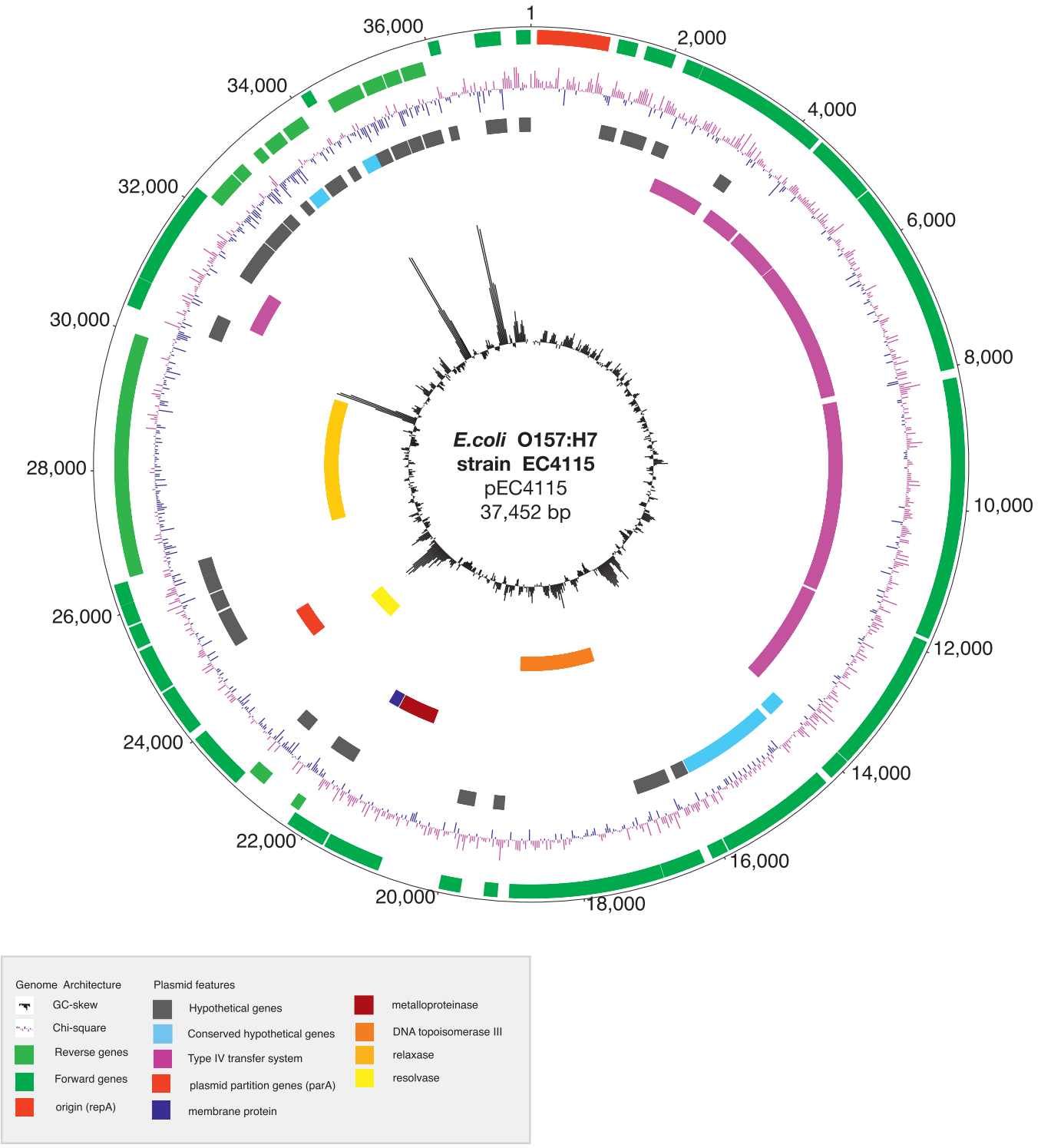

Fig. S1. Conjugal transfer plasmid pEC4115. Circles (numbered from outer to inner circle)): predicted ORFs encoded on plus strands (circle 1) (with the replication initiation gene *repA* marked in red) and on minus strands (circle 2); GC skew (circle 3); Circles 4–8 show gene inventory. The majority of genes are

Legend continued on following page

hypothetical genes with no assigned function (gray) and conserved hypothetical gens (light blue) (4). Conjugal transfer genes (magenta) (5). Membrane protein (blue), metalloproteinase (brown), and plasmid partition protein (red) (6). The DNA modification enzymes topoisomerase III (orange), relaxase (gold), and resolvase (yellow) are shown in circle 7. χ^2 is shown in circle 8.

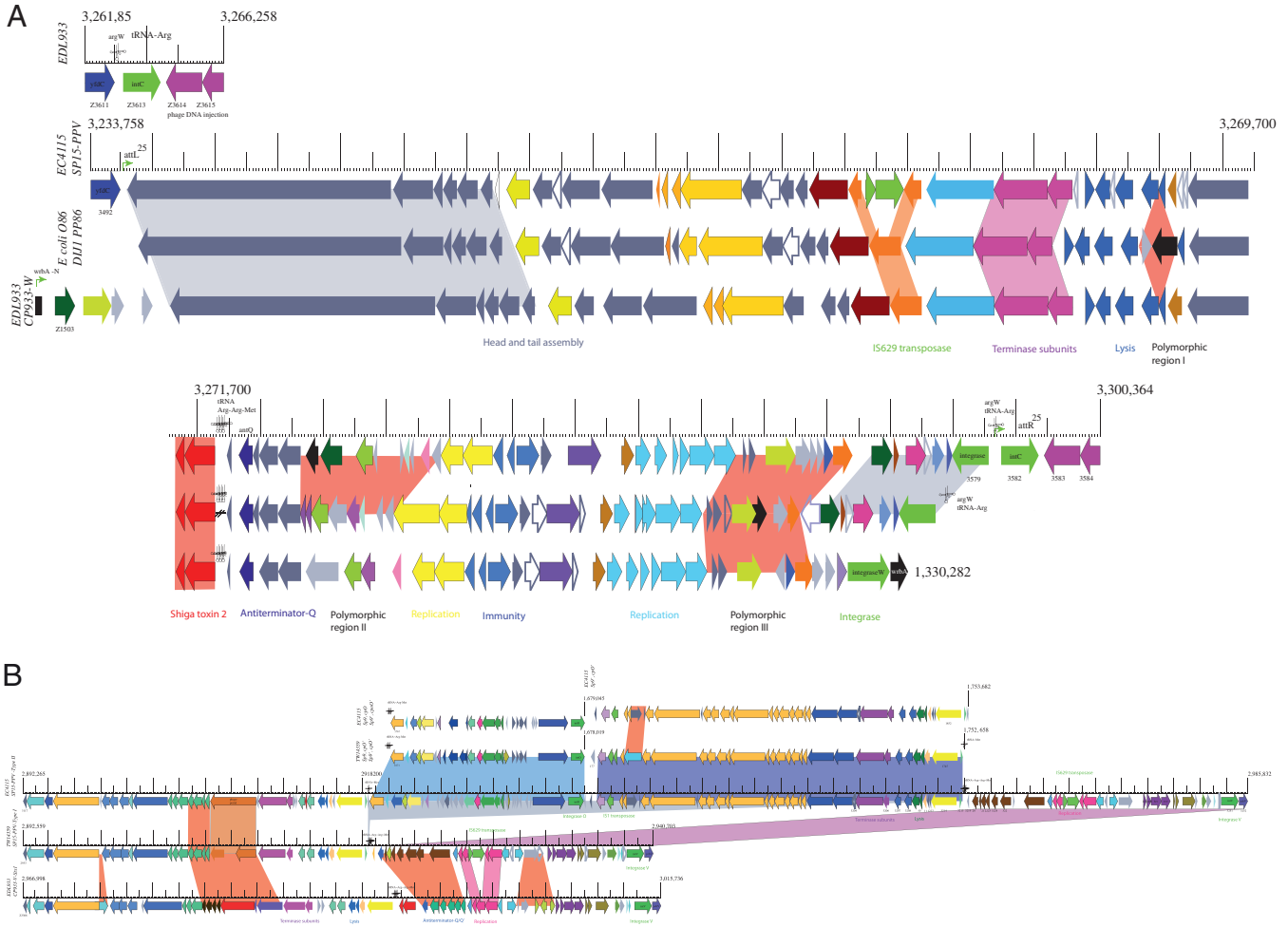

Fig. S2. Architecture and plasticity of the Stx2-converting phage and the Stx1-converting phage. (*A*) The Stx2-converting phage. The scale in bp indicates the genomic location of the prophage insertions in the respective chromosomes. The figure shows the unoccupied *argW* locus in strain EDL993 and comprehensive analyses of other Stx2-converting prophages in *E. coli*. The prophages reveal high homology and synteny in structural and enzymatic key components. This prophage type was found to occupy the *wrbA* and *argW* loci (black), and its integration typically results in perfect direct repeats (green arrows). Comprehensive analyses revealed three polymorphic regions and identified IS*629* element insertions as microevolutionary drivers. Corresponding genes are colored and shaded: hypothetical genes in light gray, conserved hypothetical genes in dark gray, gene polymorphisms caused by gene prediction in white, and mobility genes, such as transposases and phage integrases, in light green. (*B*) The Stx1-converting phage. The scale in bp indicates the genomic location of the prophage insertions in the respective chromosomes. This prophage occupies the *yehV* locus. Corresponding genes are colored and shaded: hypothetical genes in light gray, conserved hypothetical genes in dark gray, gene polymorphisms caused by gene prediction in white; and mobility genes, such as transposases and phage integrases, in light green. This figure shows two distinct prophage variants occupying the *yehV* chromosomal position in the strains associated with the SP outbreak; these variants are the key to their Stx1-negative genotype. In *stx1*-positive strains, such as EDL933, the *stx1* toxin genes are found within the cryptic prophage Sp15-PPV (48,738 bp) neighboring the tRNA triplet cluster tRNA Arg-Arg-Met. However, the SP outbreak strains show length alteration because of the lack of the *stx*1 gene operon and complex rearrangement, which accounts for an additional 46-kbp phage fragment unique to strain EC4115. The larger phage variant (93,567 bp) is unique to the ME strains, such as EC4115, whereas the smaller variant (48,144 bp) is found in the remainder of the strains associated with the SP outbreak.

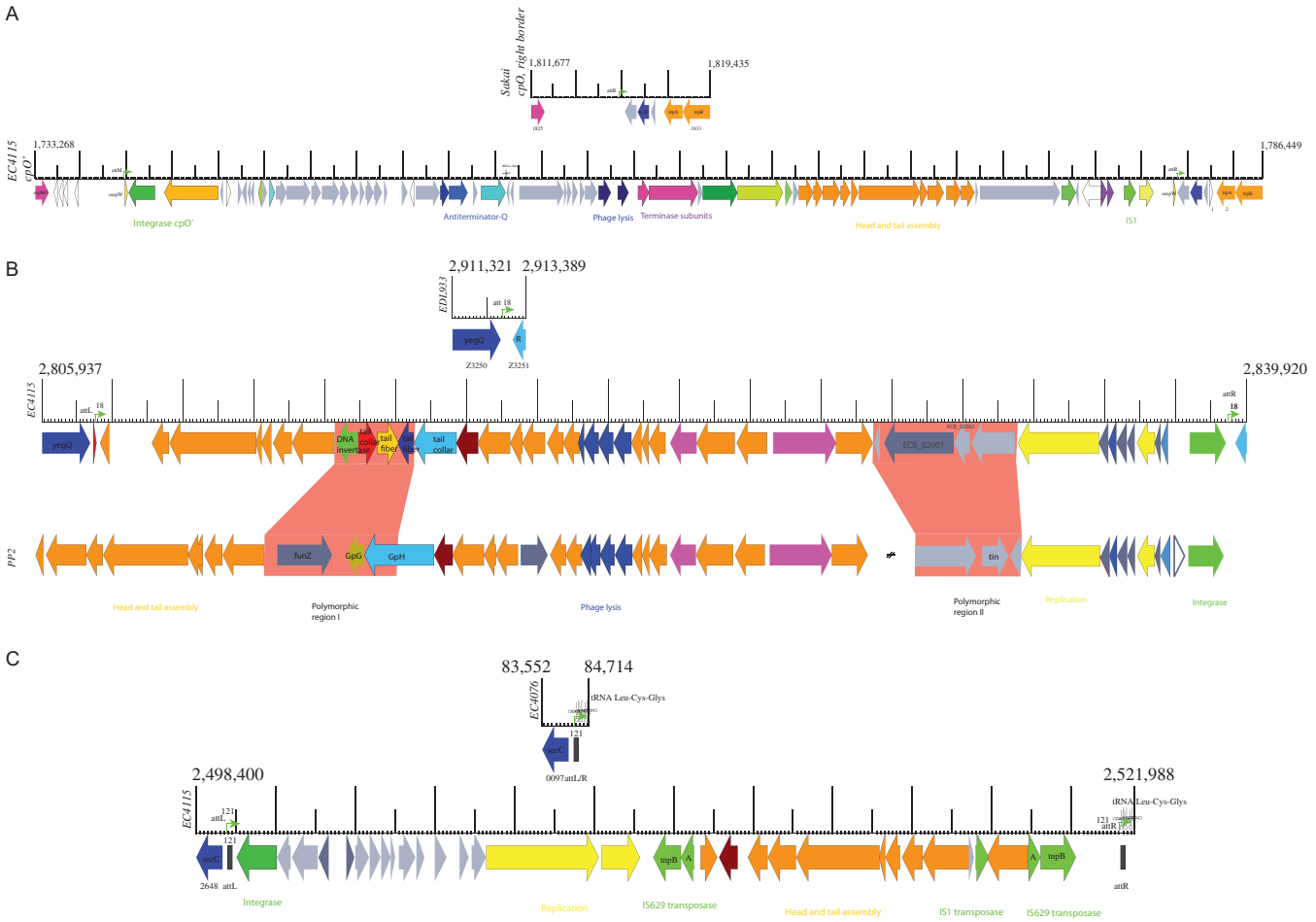

Fig. S3. (Continued)

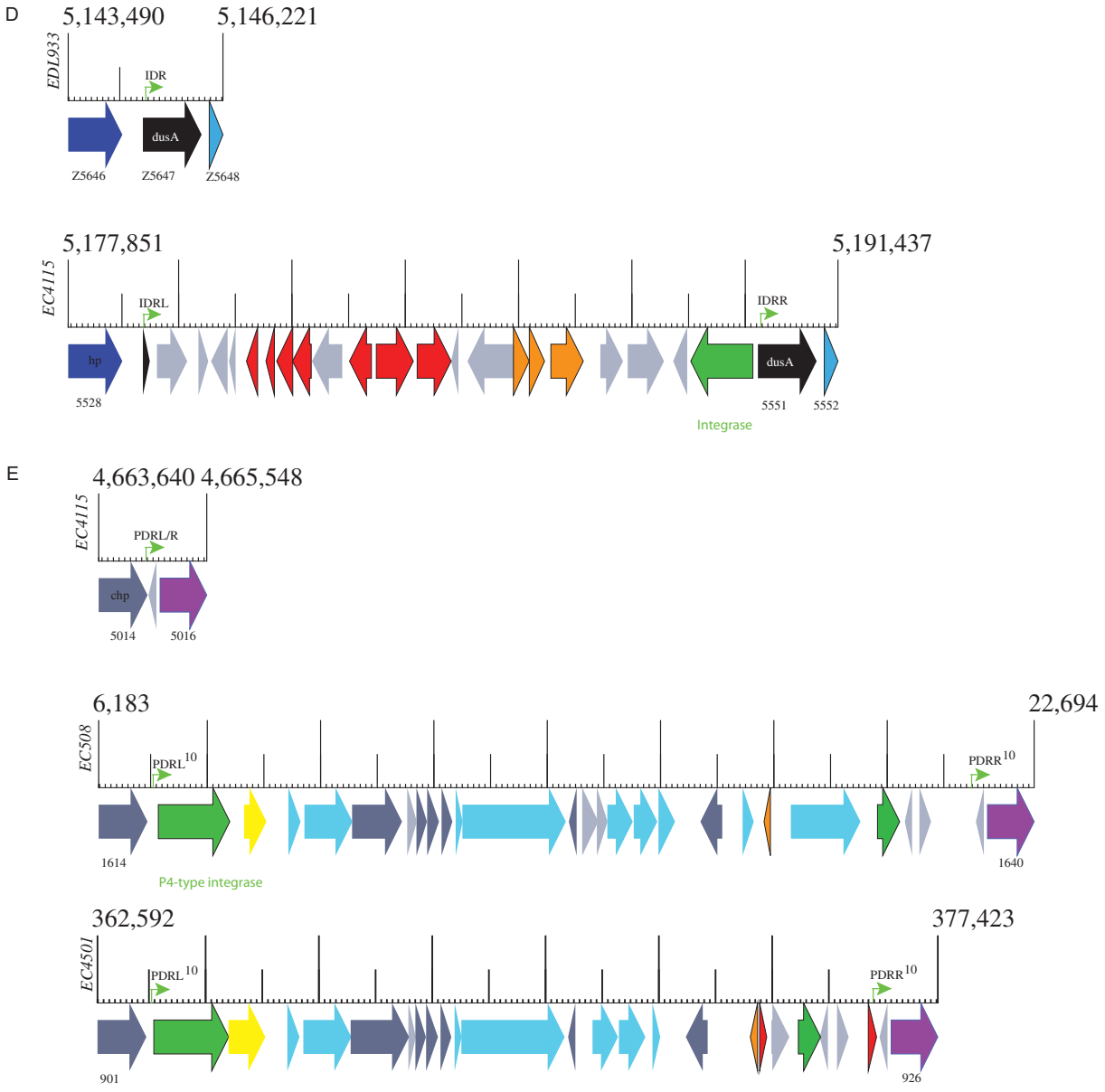

Fig. S3. (Continued)

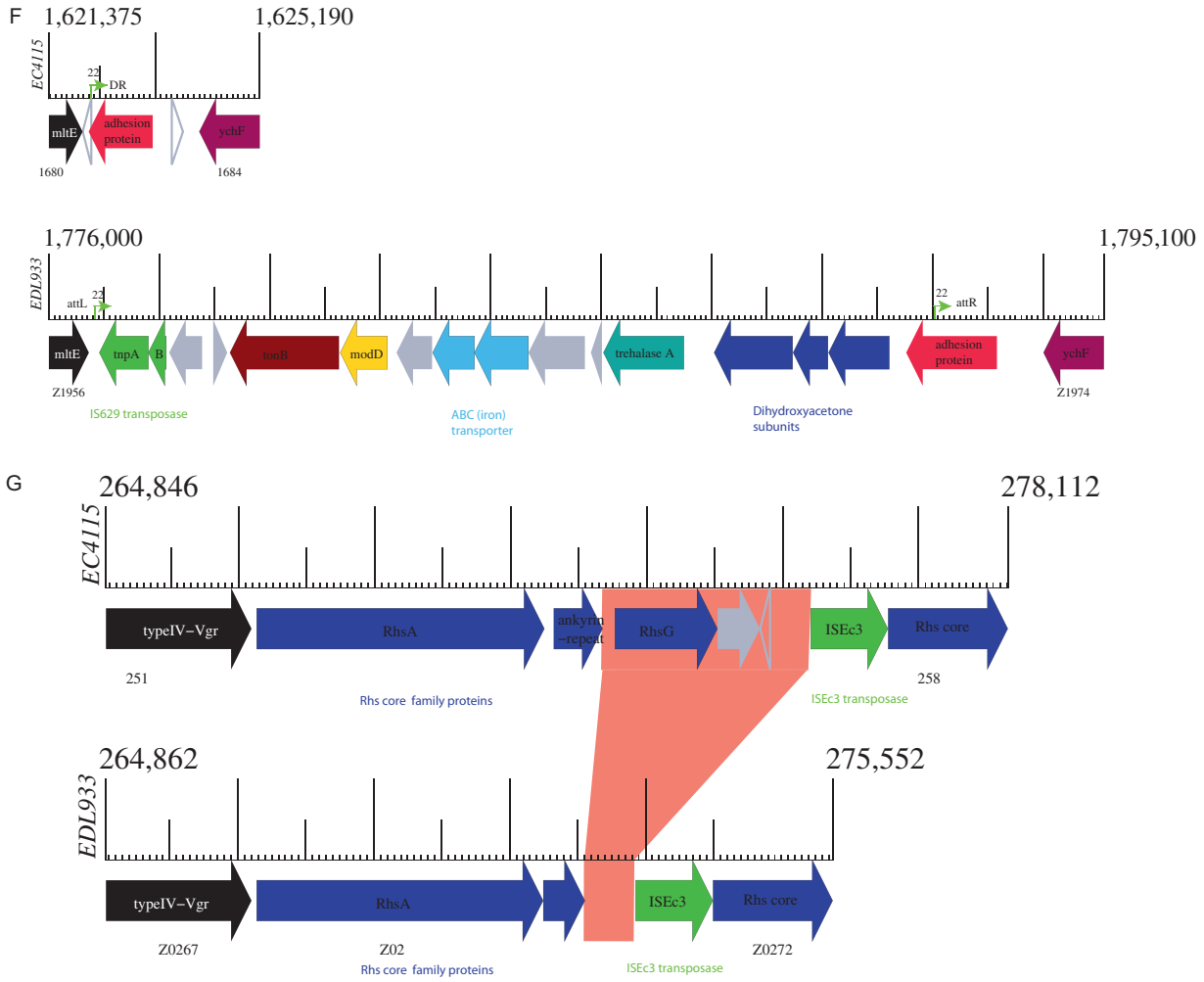

Fig. S3. Nontoxin prophage markers and remnants (*A*–*E*). The scale in bp indicates the genomic location of the prophage insertions in the respective chromosomes. Corresponding genes are colored and shaded: hypothetical genes in light gray, conserved hypothetical genes in dark gray, transposases and phage integrases in light green, and gene polymorphisms caused by gene prediction in white. Polymorphic regions are highlighted in red. Repeats caused by phage- or transposase-mediated insertions are marked as green arrows. (*A*) Lambdoid tandem prophage cpOO′. Strain EC4115 carries a lambdoid prophage insertion of 45,449 bp, cpO′, at the right border of the cpO prophage (58,298 bp) at the *ompW* locus. The complex cpOO′ phage structure features 118-bp imperfect direct repeats (Table S2) at their respective integration sides. This complex tandem phage structure of cpO and cpO′ contributes to the unique structure of the Stx1-negatiave prophage in EC4115, which is a distinguishing feature of the SP outbreak ME strains. (*B*) P2-type *Enterobacteria* phage. The 31,943-bp APEC-like prophage is unique to the SP outbreak-associated strains. This phage is integrated at the *yegQ* locus and resembles a P2-type prophage flanked by 18-bp perfect direct repeats (Table S2). The phage is organized highly syntenically and shares high protein homology with the 33,593-bp *Enterobacteria* phage P2 (NC_001895). Comprehensive analyses revealed two polymorphic regions (shaded in red) that code for structural head and tail components and a cluster of hypothetical proteins with no assigned function. The integrase is highly homologous to the *Enterobacteria* phage P2, the *E. coli* ETEC isolate O148:H28 strain B7A (ZP_03029820), and *E. coli* APEC O1:K1:H7 (1). (*C*) SP13 prophage deletion. The 21,118-bp prophage Sp13-PP is deleted in isolate EC4076. The triplet tRNA cluster tRNA Leu-Cys-Gly locus serves as the phage integration site. Integration is mediated by a phage integrase causing a 121-bp perfect direct repeat (green arrows) in typical outbreak strains, but this sequence has been reduced to a single 121-bp repeat in EC4076. (*D*) Phage scar 9.2, a 9,247-bp phage remnant, is integrated into the tRNA dihydrouridine synthase A gene (*dusA*) (ECH74115_5551) and codes 21 genes. Insertion resulted in a 104-bp imperfect direct repeat (IDRL and IDRR) (Table S2). This repeat sequence shares 84% identity with EDL933 *dusA* (Z5647). Of note, this 104-bp IDRL sequence of the N-terminal duplicated *dusA* pseudogene (ECH74115_5529) is 100% identical to the *dusA* gene in EDL933 (ECH74115_5551); however the IDRR sequences share only 84% homology. Thus, this phage integration not only inserts novel genetic material in strain EC4115 but also leads to the microevolution of this particular gene in strains EDL933 and EC4115. This phage scar codes the *Enterobacteria* phage SfV metal-dependent phosphohydrolase with conserved HD domain (2), host specificity protein J, enterobacterial Ail/Lom family protein, side-tail fiber protein, a degenerated T3SS-secreted effector NleG-homolog, and a T3SS-secreted effector NleG-like protein (3). (*E*) Phage scar 14.5, a 14,596-bp EC508 prophage, is integrated at the *dinD* locus and codes 24 genes (ECH7EC508_1615–ECH7EC508_1636). Insertion is mediated by a P4-type phage integrase and results in 10-bp perfect direct repeats. This phage codes the bacteriophage P4 integrase (ECH7EC508_1615), putative ssDNA-binding protein (ECH7EC508_1616), putative antirepressor protein (ECH7EC508_1617), and antitermination protein Q (ECH7EC508_1636). (*F*) InDel15. (*G*) InDel1.6.

1. Johnson TJ, et al. (2007) The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. *J Bacteriol* 189:3228–3236.
2. Allison GE, Angeles D, Tran-Dinh N, Verma NK (2002) Complete genomic sequence of SfV, a serotype-converting temperate bacteriophage of *Shigella flexneri*. *J Bacteriol* 184: 1974–1987.
3. Ogura Y, et al. (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci USA* 106: 17939–17944.

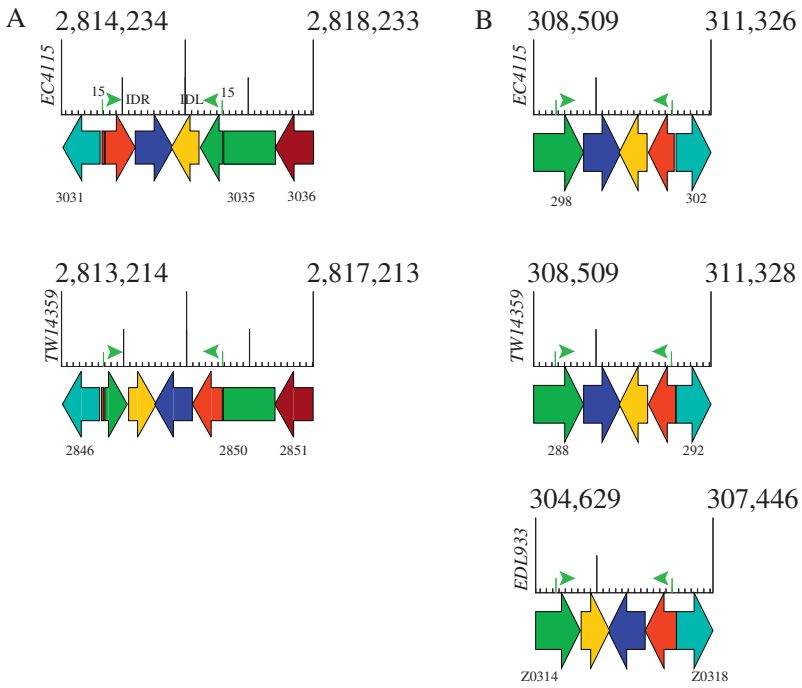

Fig. S4. Inversion locus. Inverted repeats mark inverted genomic loci within the borders of (*A*) P2-type phage and (*B*) prophage H loci. The scale in base pairs indicates the genomic location of the inversion in the respective chromosomes. Corresponding homologous genes shared between the loci are colored accordingly, and the inverted direct repeats are marked as green arrows. The resulting chimerical fusion genes code for structural phage components and are colored red and green.

Table S1. PCR primers used in this study

Name	Sequence	Minimum	Maximum	Length (bp)	Direction	Product size (bp)	Molecule
1_FP	GGCGCCTGTGGTTGTCAGAGG	397	417	21	Forward	474	pEC4115
1_RP	AGCGCCGCCTCTTTGGCTTT	851	870	20	Reverse	474	pEC4115
2_FP	AACGGCCTGAATCGGCCTGC	5,383	54,02	20	Forward	1,759	pEC4115
2_RP	TCTTCCGCAAGCTGCCACGG	7,122	71,41	20	Reverse	17,59	pEC4115
3_FP	TGGCGCGTTCAGTTGCCGAT	17,353	17,372	20	Forward	697	pEC4115
3_RP	TTGCGCTCGTCGGTGCTGAG	18,030	18,049	20	Reverse	697	pEC4115
4_FP	AGCGTCAGAACTGCGCCGAC	24,583	24,602	20	Forward	824	pEC4115
4_RP	GCCGCAAATTGCCAGGCCAC	25,387	25,406	20	Reverse	824	pEC4115
5_FP	GCAAAGCGTCACACGCGCAA	35,120	35,139	20	Forward	329	pEC4115
5_RP	GCAGGATTTGCCGGGAGCGT	35,429	35,448	20	Reverse	329	pEC4115
Inv_FP	GACGCGGTGACGGCAACAGA	1,599	1,618	20	Forward	2,260	Chromosome
Inv_RP	CCGCAAGAGCCTGTGCGGAA	3,839	3,858	20	Reverse	2,260	Chromosome

Table S2. Genome signatures of *E. coli* strain EC4115

Locus name	Locus coordinates	Genomic target	Annotation	Repeat type	Repeat coordinates	Repeat sequence	Mechanism	Length (bp)	Coding capability	Figure
InDel 1.6	275,212	ECH74115_0251	Ankyrin repeat gene	Not detected	Not applicable		*ISEc3* transposase (ECH74115_0257)	1,667	Rhs core gene cluster	Fig. S6*B*
Inversion loci I & II	308,846	ECH74115_3031 to _3035, ECH74115_298 to _302	Phage tail fiber proteins	Inverted direct repeats	15 p (308,846–308,860; 310,705–310,691; 2,814,895–2,814,909; 2,816,799–2,816,785)	ttggtttgggagaag - cttctcccaaaccaa	Homologous recombination	1,9	Structural phage genes	Fig. S7*A* (locus I) Fig. S7*B* (locus II)
InDel 15	1,622,120	ECH74115_1682	Adhesion protein	Imperfect direct repeat	25 bp (1,622,120–1,622,141)	**Hybrid:**ATTCTtgCgc cgctgAgCGCCC **IDRL:**ATTCTtgCcat ctgaAcCGCCC **IDRR:** ATTCTacCgc cgctgAgCGCCC Bases that were the same in both 22-bp sequences on the EDL933 chromosome are capitalized	*IS629* transposase (Z1957, Z1958)	15,147	Iron ABC transporter, dihyroxy-acetone, adhesion protein	Fig. S6*A*
cpOO′ tandem prophage	1,737,154	*ompW (ECH74115_1820, cpO′, ECH74115_1887, cpO′)*	*ompW* locus	Imperfect direct repeat	117 bp (1,737,154–1,737,270; 1,782,576–1,782,692)	**IDRL:**gtgtggtacatgg atatcgataccacggcta agtataaatctggtgtta ctaccgtgaaagactcg gtacgactggatccgtg ggtgtttatgttctccgc aggatatcgtttttaa **IDRR:** gtgtggtacatg gatatcgataccaccgcc aattataaactgggcgg tgc-acagcaacatgata gcgtacgcctcgatccgt gggtgtttatgttctcagc aggatatcgtttttaa	Phage integrase cpO′ (ECH74115_1821)	45,306	cpO(O′) phage genes	Fig. S4*A*, Fig. S3
SP13-deletion in strain EC4076	2,499,194	ECH74115_2581 to _2683	tRNA Leu-Cys-Glys	Perfect direct repeat	121 bp (2,499,194–2,499,314; 2,521,631–2,521,750)	aaaaaaaccacccgaa ggtggtttcacgacactg cttattgctttgattttat tcttatctttcccatggtac ccggagcgggacttga acccgcacagcgcgaa cgccgagggatttttaa	Phage integrase (ECH74115_2649)	21,118	Phage - associated genes, I*S629* and *IS1*	Fig. S4*C*
Stx2c prophage	2,662,922	*sbcB* (ECH74115_2944)	*sbcB* locus	Perfect direct repeats	EC4115, 13 bp: (2,662,922–2,662,934; 2,720,134–2,720,146) (PP1717, 13 bp; YYZ2008, 21 bp)	**EC4115:** tttcacgat tacg **P1717:** tttcacgat tacg **YYZ2008:** cacgcgcg taacgtgacaggg	Phage integrase **intX** (ECH74115_2942)	57,224	Phage-associated genes, *IS629* and *IS66*	Fig. 3

Table S2. Cont.

Locus name	Locus coordinates	Genomic target	Annotation	Repeat type	Repeat coordinates	Repeat sequence	Mechanism	Length (bp)	Coding capability	Figure
APEC-like phage	2,807,452	ECH74115_3021	*yegQ* peptidase U32 family protein	Perfect direct repeat	18 bp (2,807,452–2,807,469; 2,839,412–2,839,429)	tacacgggcttatttttt	APEC-like phage integrase (ECH74115_3066)	31,943	Phage-associated genes	Fig. S4*B*
ME strain-specific Stx1 phage insertion	2,918,717	tRNA (Arg-Arg)-Met	**ECH74115_3149 (check)** adjacent to tRNA Arg-Arg-Met	Perfect direct repeat	tRNA Met repeat **(**2,918,717–2,918,643)	gccctttagctcagtggtgagagcgagcgacgcataatcgccaggtcgctggttcgaatccagtacaacg cgcca	Phage integrase *intO* (ECSP_1651, ECH74115_3178)	45,632	Phage- associated genes	Fig. S3, Fig. S4*A*
Stx1 negative phage	2,985,184	*yehV* (ECH74115_3252)	*yehV* locus	Not detected	Not applicable		Phage integrase *intV* (ECH74115_3251)	93,567	Phage-associated genes	Fig. S3
Stx2 prophage	3,234,765	*argW* (ECH74115_3581)	tRNA Arg adjacent to phage integrase and two phage related DNA injection proteins (ECH74115_3582 - ECH741153584)	Perfect direct repeat	25 bp (3,234,765–3,234,789; 3,297,023–3,297,047).	tgtcctcttagttaaatggatataa	Phage integrase (ECH74115_3579)	62,282	Phage -associated genes	Fig. S2
Phage scar 14.5 (EC508)	4,664,723	ECH74115_5016	DNA damage inducible protein	Perfect direct repeat	10 bp, (EC508, NZ_ABHW01000007: 7023–7033; 21628–21638)	gagcagattc	P4 type phage integrase (ECH7EC508_1615)	14,596	Phage -associated genes, phage remnant	Fig. S5*B*
Phage scar 11 (EC869)	4,683,872	ECH74115_5031	tRNA Sec neighbored by remnant phage integrase and *IS91* transposase (ECH74115_5032 to ECH74115_5034)	Imperfect direct repeats	30 bp (EC869, NZ_ABHU01000002: 402,713–402,742; 413,782–413,811;)	**IDRL:** ggcaggttcgactcctgtgatcttccgcca **IDRR:** ggcaattttgactcctgtgatcttccgcca	P4 phage integrase (ECH7869_1671)	11,040	Phage -associated genes, phage remnant	Fig. 1
Phage scar 9.2	5,179,169	ECH74115_5551, ECH74115_5529	tRNA ihydrouridine synthase *dusA*	Imperfect direct repeat	104 bp (5,179,169–5,179,272; 5,190,024–5,190,127)	**IDRL:**atgcacggtaattctgaaatgcaaaaaatcaaccaaaccagcgcaatgcctgaaaaaactgacgttcactggagtggtcggtttagcgttgcaccaatgctcga **IDRR:**atgcacgataatcatgaagcacaaaaaattaatcaaaccagcgtgatgcctgaaaaactggtgtttactggaattctcggtttagcgttgcaccaatgctcga	phage integrase (ECH74115_5550)	10,750	phage associated genes, phage remnant	**Fig. S5*A***

Dataset S1. Strain history and associated metadata

Dataset S1

Dataset S2. SNP discovery in the *E. coli* O157:H7 lineage

Dataset S2

This dataset lists 1,225 intra- and intergenic identified SNPs for the *E. coli* O157:H7 lineage. Functional annotation of intragenic SNPs and resulting transition or transversion for each SNP are listed. SNP positions are referenced to the respective location on the strain EC4115 chromosome. Nucleotide bases colored in blue (reference) and red were validated in the respective strains using pyrosequencing assays. B, bovine; gt, genotyped; RSC, reference strain collection; SO, spinach outbreak; TB, Taco Bell outbreak; TJ, Taco John outbreak.

Dataset S3. Pyrosequencing assay. Population genetic analyses of 229 strains using 19 canonical SNPs

Dataset S3