Ancestral Divergence, Genome Diversification, and Phylogeographic Variation in Subpopulations of Sorbitol-Negative, β -Glucuronidase-Negative Enterohemorrhagic *Escherichia coli* O157†

JAEHYOUNG KIM,¹ JOSEPH NIETFELDT,¹ JINGLIANG JU,¹ JOHN WISE,¹ NARELLE FEGAN,²
PATRICIA DESMARCHELIER,² AND ANDREW K. BENSON^{1*}

Department of Food Science and Technology, University of Nebraska, Lincoln, Nebraska 68583-0919,¹ and Food Safety and Quality Group, Food Science Australia, Cannon Hill, Queensland, Australia, 4170²

Received 21 May 2001/Accepted 4 September 2001

The O157:H7 lineage of enterohemorrhagic *Escherichia coli* is a geographically disseminated complex of highly related genotypes that share common ancestry. The common clone that is found worldwide carries several markers of events in its evolution, including markers for acquisition of virulence genes and loss of physiological characteristics, such as sorbitol fermentation ability and β -glucuronidase production. Populations of variants that are distinct with respect to motility and the sorbitol and β -glucuronidase markers appear to have diverged at several points along the inferred evolutionary pathway. In addition to these variants, distinct subpopulations of the contemporary non-sorbitol-fermenting, β -glucuronidase-negative O157:H7 clone were recently detected among bovine and human clinical isolates in the United States by using high-resolution genome comparison. In order to determine if these recently described subpopulations were derived from a regional or ancestral divergence event, we used octamer-based genome scanning, marker sorting, and DNA sequence analysis to examine their phylogenetic relationship to populations of non-sorbitol-fermenting, β -glucuronidase negative O157:H7 and O157:H– strains from Australia. The inferred phylogeny is consistent with the hypothesis that subpopulations on each continent resulted from geographic spread of an ancestral divergence event and subsequent expansion of distinct subpopulations. Marker sorting and DNA sequence analyses identified sets of monophyletic markers consistent with the pattern of divergence and demonstrated that phylogeographic variation occurred through emergence of regional subclones and concentration of regional polymorphisms among distinct subpopulations. DNA sequence analysis of representative polyphyletic markers showed that genome diversity accrued through random drift and bacteriophage-mediated events.

Hemorrhagic colitis is caused by a number of serotypes of Shiga toxin-producing *Escherichia coli* (STEC) (14). Among the clinical STEC strains that have been isolated, a subset of enterohemorrhagic *E. coli* (EHEC) strains has been found which carry common sets of virulence genes that encode factors for attachment to host cells, elaboration of effector molecules, and production of two different types of Shiga toxins (22). The sets of virulence genes are found in the locus of enterocyte effacement (LEE) pathogenicity island, lambdoid bacteriophages, and a large virulence-associated plasmid (8, 9, 23, 25, 26, 31, 32). Population genetic analysis of EHEC and STEC strains has shown that EHEC strains comprise two divergent lineages, termed EHEC 1 and EHEC 2, that are only distantly related but apparently experienced similar pathways of virulence gene acquisition (24, 28, 38). The EHEC 1 lineage is comprised solely of a geographically disseminated cluster of strains with related genotypes bearing O157:H7 and O157:H– serotypes, while the EHEC 2 lineage is serotypically and genotypically more diverse.

The O157 serotype can be found in genetically diverse populations of *E. coli*, apparently as a consequence of transfer of

a common allele combination in the *gnd-rfb* region that encodes the O157 antigen (36). Among the populations bearing the O157 serotype, only the genetically related populations in the serotype O157:H7 EHEC 1 lineage, including the derived nonmotile O157:H– populations (collectively referred to as O157 EHEC), carry the combination of virulence genes that includes the genes found in the LEE island, the large virulence plasmid, and the Shiga toxin-converting phages (38). A reconstruction of the evolutionary history of this lineage suggests that it arose from transfer of the O157 *gnd-rfb* region into an O55:H7 enteropathogenic *E. coli* host bearing the LEE island and an *stx2*-converting prophage (11, 28, 36, 38, 39). In subsequent steps, this ancestral population lost the ability to ferment sorbitol, was lysogenized by an *stx1*-converting phage, and finally acquired a mutation that inactivated the *uidA* gene, which resulted in a loss of β -glucuronidase activity (11, 39).

The contemporary non-sorbitol-fermenting, β -glucuronidase-negative O157:H7 clone is the EHEC that is most frequently isolated from hemorrhagic colitis patients in the United States, Canada, Japan, and the United Kingdom. Significant numbers of nonmotile O157 EHEC strains have been recovered from samples from continental Europe and Australia (6, 15, 19, 29). The nonmotile strains found in continental Europe are sorbitol positive and β -glucuronidase positive, and phylogenetic analysis suggests that they comprise a population that diverged from an ancestral intermediate of the contemporary O157:H7 clone prior to loss of the sorbitol fermentation

* Corresponding author. Mailing address: Department of Food Science and Technology, University of Nebraska, 330 Food Industry Complex, Lincoln, NE 68583-0919. Phone: (402) 472-5637. Fax: (402) 472-1693. E-mail: abenson1@unl.edu.

† Journal Series paper 13460 of the Nebraska Agricultural Research Experimental Station.

and β -glucuronidase production traits (11, 19). In contrast, nonmotile O157 EHEC strains that are commonly isolated in Australia are sorbitol negative and β -glucuronidase negative (29), suggesting that they are members of a more recently derived population that diverged from the lineage after loss of the sorbitol fermentation and β -glucuronidase production traits.

In addition to populations of nonmotile variants, two distinct subpopulations of the widespread non-sorbitol-fermenting, β -glucuronidase-negative O157:H7 clone have been detected in the United States by high-resolution genome comparison (20). Strains belonging to these two subpopulations, termed lineages I and II, appeared to be nonrandomly distributed among the bovine and human clinical isolates examined and could be distinguished by octamer-based genome scanning (OBGS) and by restriction fragment length polymorphism analysis with lambdoid phage probes. Since these studies were conducted exclusively with United States strains, it was not clear whether one or both of the subpopulations are regional subclones or whether they resulted from a more ancestral divergence event.

To determine the ancestry of the lineage I and II subpopulations, we studied the phylogenetic relationships of representative lineage I and II O157:H7 strains from the United States to O157 EHEC strains from the geographically isolated continent Australia. The phylogeny inferred from cladistic and statistical analyses of OBGS data shows that populations of O157 EHEC strains isolated from the two continents comprise two lineages corresponding to the distribution of lineage I and II United States strains, indicating that the divergence of the two lineages was ancestral and predated the arrival of the lineages on one or both continents. Results obtained by sorting polymorphic OBGS markers relative to the inferred phylogeny of the strains are consistent with this conclusion and were used to identify sets of monophyletic markers that distinguished divergence of the lineages and emergence of three independent lineage II subpopulations. Two mechanisms of phylogeographic variation were also detected, and these mechanisms corresponded to expansion of regional subclones and accumulation of regional polymorphisms in a lineage-independent fashion. DNA sequence analysis of monophyletic markers provided a set of specific genome alterations that marked divergence of the lineages and subpopulations. Sequencing of polyphyletic markers suggested that random drift in the populations is frequently a result of common alterations that randomly create or destroy OBGS priming sites and events associated with movement and recombination among lambdoid prophages.

MATERIALS AND METHODS

Bacterial strains and growth conditions. Strains used in this study and their known characteristics are shown in Table 1. United States isolates have been described previously (10, 13, 20, 21, 33) and were chosen because they represent the genetic diversity of the strains in the United States that have been examined by OBGS analysis. Strains from Australia and New Zealand were chosen on the basis of temporal and spatial representation of strain collections. The set of 15 human isolates included 2 isolates from patients with hemolytic-uremic syndrome in New Zealand and 13 clinical isolates from patients with diarrhea or hemolytic-uremic syndrome in Australia (kindly supplied by R. Robins-Browne and D. Lightfoot, University of Melbourne, Parkville, Victoria, Australia). Isolates of animal origin were obtained from bovine or ovine sources in Australia during various pre- and postharvest surveys of animals. All strains were main-

tained as frozen stock preparations and were minimally propagated on Luria-Bertani agar or broth.

OBGS analysis. DNA was prepared from 10-ml cultures of each strain by standard methods, and 50 ng was used for each OBGS reaction. Reactions were primed with combinations of decamer primers, each of which had a 5' AT dinucleotide followed by a skewed octamer sequence (20). The 5' AT dinucleotide, which was added to limit steric effects of the 5' fluorophore on annealing of the octamers, was also included in the sequences of the unlabeled octamers. For each reaction, a labeled primer (containing a leading strand-biased octamer) was combined with a cold primer (containing a lagging strand-biased octamer). The following primers were labeled at their 5' ends with the fluorophore IRD800 (Li-Cor, Inc., Lincoln, Nebr.): OCT3B (ATGCTGGTGG), OCT4B (ATGCTG GCGG), OCT21 (ATGCGCTGGA), and OCT22 (ATCTGCGCAA). The following unlabeled primers were used with the fluorescent primers: OCT4C (AT CCGCCAGC), OCT6C (ATGCCAGCGC), OCT12C (ATGCCGCCAG), OCT13C (ATTGCGCCAG), and OCT19C (ATCTTCCAGC). The combinations OCT3B-OCT4C, OCT3B-OCT12C, OCT4B-OCT19C, OCT21-OCT13C, and OCT22-OCT6C were used for OBGS studies. The fluorescent OBGS products were resolved by fragment analysis with Li-Cor 4200 automated DNA sequencers, and data were collected with the Global IR² system (Li-Cor, Inc.). Strain MC1061, a K-12 derivative, and a distantly related bovine STEC strain, AU133c, were included in the analyses to assess the ancestral condition of the markers. Band patterns were printed with an Alden 9315 thermal printer and scored as described previously (20).

OBGS data analysis. Binary strings, representing the presence or absence of OBGS products ranging from 200 to 1,500 bases long for each strain, were constructed in Microsoft Excel spreadsheets from gel images of products obtained with individual OBGS primer combinations. The binary strings for each of the primer combinations were then added head to tail for each strain with FORMATALL, a Perl-based program that creates composite files from fragment analysis and microarray data files and generates output files compatible with PAUP and other clustering programs (Wise and Benson, unpublished data). Genetic distance was calculated by determining the mean number of character differences from all pairwise alignments of taxa, and phylogenetic trees were resolved by neighbor joining (NJ) and statistical (bootstrap) sampling of the data using algorithms in PAUP, version 4.0 (34). Trees generated from binary strings resulting from individual primer combinations were in general agreement with the tree generated from the composite binary strings resulting from all primer combinations by both NJ and bootstrap methods.

Polymorphic OBGS markers were sorted by using macro functions in Microsoft Excel spreadsheets. The binary data were first rendered in columns for each strain by the FORMATALL program and output into a single Excel spreadsheet. The data were sorted in phylogenetic order by ranking the columns of binary data for each strain according to the position of the strain in the phylogenetic tree. When the columns were sorted in this order, polymorphic characters exhibited one of four general patterns of distribution relative to phylogeny; class I characters were lineage specific and conserved, class II characters were lineage specific and not conserved but were clustered in monophyletic groups, class III characters were lineage specific and not conserved but were distributed in polyphyletic groups, and class IV characters were lineage independent and distributed in polyphyletic groups. In sequential steps, the macro functions identified polymorphic characters, classified them in the four general patterns of distribution, rendered them as cells with different colors by conditional formatting, and grouped the polymorphisms belonging to each class for enumeration. A similar set of macro functions was iterated to identify geographical markers, except that the strains were sorted primarily by geography and were grouped secondarily on the basis of phylogeny and distribution class. The macros are available from us upon request.

OBGS band isolation and sequencing. Polymorphic OBGS products that indicated divergence of the lineages and clades were identified from images of gels in which the samples were loaded in phylogenetic order. Representative bands for each class were excised from automated sequencing gels in which electrophoresis was halted immediately after elution of the appropriate molecular weight markers past the scanning optical unit of the sequencer. The bands of interest were then localized in a gel by drying the gel on filter paper and scanning the dried gel with an Odyssey near-infrared fluorescence scanner (Li-Cor, Inc.). After localization and excision, fluorescently labeled DNA was eluted from a gel slice by boiling for 5 min in water, followed by phenol extraction and ethanol precipitation. The precipitated pellet was redissolved in 50 μ l of water, and 10 μ l was used as a template in a 20- μ l OBGS reaction mixture containing the primer combination and reagents that were present in the original mother reaction mixture. The reamplification products were then electrophoresed alongside the mother reaction mixtures to check band purity and efficiency of reamplification.

TABLE 1. Strains used in this study

Strain	Year of isolation	Source	Sample or disease ^a	Production	Geographical location	Serotype	Stx ^b	eae ^c	ehxA ^c
AU6	1995	Calf	Feces	Dairy	Southeast Queensland, Australia	O157:H7	Stx2	+	+
AU14	1995	Lamb	Chop	Meat	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU119	1996	Cattle	Carcass	Meat	Southeast Queensland, Australia	O157:H7	Stx2	+	+
AU134	1996	Cattle	Feces	Meat	Southeast Queensland, Australia	O157:H7	Stx2	+	+
AU151b	1996	Cattle	Feces	Meat	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU183	1996	Sheep	Feces	Meat	New South Wales, Australia	O157:H-	Stx1, Stx2	+	+
AU195c	1996	Calf	Feces	Dairy	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU197b	1996	Calf	Feces	Dairy	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU200	1996	Cattle	Carcass	Meat	Southeast Queensland, Australia	O157:H7	Stx1, Stx2	+	+
AU329	1996	Cattle	Feces	Meat	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU514	1993	Cattle	Boned beef	Meat	Victoria, Australia	O157:H7	Stx2	+	+
AU516	1993	Calf	Feces	Dairy	Northern Territory, Australia	O157:H-	Stx1, Stx2	+	+
AU525	1993	Calf	Carcass	Dairy	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU539	1993	Sheep	Feces	Meat	Southeast Queensland, Australia	O157:H-	Stx1	+	+
AU571	1993	Calf	Feces	Dairy	Southeast Queensland, Australia	O157:H-	Stx1, Stx2	+	+
AU580a	1994	Sheep	Carcass	Meat	Western Australia, Australia	O157:H-	Stx1, Stx2	+	+
AU581	1994	Cattle	Carcass	Meat	South Australia, Australia	O157:H-	Stx1, Stx2	+	+
AU623	1997	Cattle	Hides	Meat	Southeast Queensland, Australia	O157:H7	Stx2	+	+
AU726	1997	Cattle	Feces	Dairy	Southeast Queensland, Australia	O157:H7	Stx2	+	-
AU735	1997	Cattle	Feces	Dairy	Far North Queensland, Australia	O157:H7	Stx2	+	+
AU1180	1998	Cattle	Feces	Dairy	Southeast Queensland, Australia	O157:H7	Stx1, Stx2	+	+
AU1668	1999	Cattle	Carcass	Meat	Central Queensland, Australia	O157:H-	Stx1, Stx2	+	+
FRIK1641	1995	Raccoon	Feces		Wisconsin	O157:H7	Stx2	+	+
NE1487	1998	Cattle	Feces	Meat	Nebraska	O157:H7	Stx1, Stx2	+	+
FRIK1574	1995	Cattle	Feces	Dairy	Wisconsin	O157:H7	Stx1, Stx2	+	+
FRIK920	1995	Cattle	Feces	Dairy	Wisconsin	O157:H7	Stx1, Stx2	+	+
FRIK1123	1995	Cattle	Feces	Dairy	Wisconsin	O157:H7	Stx1, Stx2	+	+
FRIK1996	1991	Cattle	Feces	Dairy	Maryland	O157:H7	Stx1, Stx2	+	+
FRIK1999	1991	Cattle	Feces	Dairy	Nebraska	O157:H7	Stx1, Stx2	+	+
FRIK1988	1991	Cattle	Feces	Dairy	Idaho	O157:H7	Stx2	+	+
AU1808	1991	Human	Diarrhea		New South Wales, Australia	O157:H-	Stx1	+	+
AU1809	1991	Human	Diarrhea		New South Wales, Australia	O157:H-	Stx2	+	+
NZ1810	1993	Human	Hemolytic-uremic syndrome		New Zealand	O157:H7	Stx2	+	+
NZ1811	1993	Human	Hemolytic-uremic syndrome		New Zealand	O157:H7	Stx2	+	+
AU1812	1996	Human	Bloody diarrhea		Southeast Queensland, Australia	O157:H7	Stx1, Stx2	+	+
AU1814	1988	Human	Bloody diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1815	1997	Human	Bloody diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1816	1999	Human	Bloody diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1817	1999	Human	Diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1818	1986	Human	No diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1819	1999	Human	Bloody diarrhea		Victoria, Australia	O157:H-	Stx1, Stx2	+	+
AU1820	1986	Human	Hemolytic-uremic syndrome		Victoria, Australia	O157:H7	Stx2	+	+
AU1821	1995	Human	Hemolytic-uremic syndrome		Victoria, Australia	O157:H-	Stx1, Stx2	+	-
AU1822	1996	Human	Hemolytic-uremic syndrome		Victoria, Australia	O157:H-	Stx2	+	+
AU1823	1987	Human	Unknown		Victoria, Australia	O157:H-	Stx1	+	+
FRIK523	1994	Human	Bloody diarrhea		Wisconsin	O157:H7	Stx1, Stx2	+	+
FRIK535	1994	Human	Bloody diarrhea		Wisconsin	O157:H7	Stx1, Stx2	+	+
FRIK585	1994	Human	Bloody diarrhea		Wisconsin	O157:H7	Stx1, Stx2	+	+
93-001	1993	Human	Bloody diarrhea		Washington	O157:H7	Stx1, Stx2	+	+
95-003	1995	Human	Bloody diarrhea		Washington	O157:H7	Stx1, Stx2	+	+
EDL933	1982	Cattle	Hamburger		Michigan	O157:H7	Stx1, Stx2	+	+
FDA518	1982	Human	Unknown		California	O157:H7	Stx1, Stx2	+	+
FDA520	1982	Human	Unknown		Oregon	O157:H7	Stx1, Stx2	+	+

^a Clinical characteristics.^b Determined by hybridization or PCR.^c Determined by PCR.

The reamplified products were then used as templates in a second round of reamplification performed with the following OBGS primers tailed with M13 forward (M13F) and reverse (M13R) primer sequences: M13FOCT3B (CACG ACGTTGTAAAACGACATGCTGGTGG), M13FOCT21 (CACGACGTTGT

AAAACGACATGCGCTGGA), M13FOCT22 (CACGACGTTGTAAAACGA CATCTGCGCAA), M13ROCT4C (GGATAACAATTTACACAGGGCCGC CAGCT), M13ROCT6C (GGATAACAATTTACACAGGATGCCAGCGC), M13ROCT12C (GGATAACAATTTACACAGGATGCCGCCAG), and

M13ROCT13C (GGATAACAATTTACACAGGATTGCGCCAG). Tailed PCR products were purified from agarose gels with a QIAquick gel extraction kit (Qiagen, Inc., Valencia, Calif.). Purified products were cloned into pTOPO (Invitrogen, Carlsbad, Calif.) and sequenced with T7 and T3 sequencing primers. DNA sequence analysis was performed for both strands.

Nucleotide sequence accession numbers. The GenBank accession numbers for the DNA sequences determined in this study are AF368044 (226C-3), AF368045 (226C-4), AF368046 (3B4C-5), AF368047 (226C-13), AF368048 (226C-14), AF368049 (2212C-1), AF368050 (3B12C-2), AF368051 (3B4C-2), AF368052 (2113C-1), and AY036017 (3B4C-1).

RESULTS

OBGS analysis of United States and Australian O157:H7 isolates. To examine the genetic relationships of Australian O157 EHEC strains and United States lineage I and II O157:H7 strains, we used OBGS to analyze a set of 37 isolates obtained from cattle and humans in Australia and eight representative United States isolates belonging to each of the two lineages (Table 1). The Australian strains were chosen because they represented broad ranges of temporal and spatial diversity with respect to origin and source. The United States isolates belonging to lineages I and II represented the spectrum of genetic diversity observed for strains derived from these lineages (20). Five different OBGS primer combinations were used for each strain, and OBGS products obtained with each primer combination were converted to binary characters. Binary strings for each primer combination were added head to tail, which resulted in composite binary strings consisting of 1,159 characters for each strain. Strain AU133c, a bovine-derived STEC strain that is distantly related to O157:H7 based on the distribution of OBGS polymorphisms, was included in the strain set as an outgroup.

Of the 1,159 characters scored, 258 (23%) were variable in the O157 strains tested, indicating that significant diversity was detected in the strain set. A total of 163 of these polymorphic characters were informative. As shown in Fig. 1, NJ analysis resolved two main clusters corresponding to the distribution of lineage I and II O157:H7 strains obtained from the United States. Bootstrap values indicated that assignment of the strains to these lineages was statistically significant. Australian O157:H7 and O157:H[−] EHEC isolates occurred in both lineages, and the majority of the strains (31 of the 37 strains) exhibited higher levels of genetic relatedness to the lineage II United States isolates.

Based on a 50% cutoff for significance, each of the major clusters was significant, suggesting that the two lineages could be divided into five clades (clades A to E) that were segregated primarily by lineage and secondarily by geography. This pattern of inferred phylogeny suggests that the two lineages previously detected in the United States (20) are not regional subpopulations but instead are descendants resulting from an ancestral divergence event that preceded geographic dissemination. The presence of the H[−] character state in most strains in clade B and all strains in clade D implies that loss of motility is associated with divergence of these clades from their lineage I and II ancestors.

Genetic markers and phylogeny of United States and Australian O157:H7 and O157:H[−] EHEC strains. If the clades inferred from the phylogenetic analysis indeed constitute distinct subpopulations, then signatures that indicate their phylogeny and diversification should be present and observable in

polymorphic OBGS products. To mine the putative signatures from the OBGS data, we developed a marker-sorting method that identified and quantified polymorphisms with monophyletic and polyphyletic distributions. This method required classification of polymorphisms into four classes, classes I to IV, based on four possible patterns of distribution relative to user-specified groups. Relative to phylogeny, the four distribution classes are shown in Fig. 2; to obtain the data, the OBGS samples were loaded in inferred phylogenetic order on the gel.

The first three classes of polymorphisms, classes I to III, occurred exclusively in an inferred lineage. Class I bands were conserved in all members of the lineage. If the predicted phylogeny is valid, the exclusiveness and conservation of class I bands in all members of a lineage imply that the bands were derived from ancestral genome alterations fixed by periodic selection during divergence of two lineages. Class II bands were conserved in monophyletic groups of strains in a lineage, and their distribution relative to the inferred phylogeny implies that they arose from genome alterations that occurred after lineages diverged. The class III bands were lineage specific but were distributed among polyphyletic groups in a lineage. The simplest explanations for the distribution of these bands are that they originated from genetic drift among alterations ancestral to the lineage and that they originated by lateral spread of common alleles in members of a lineage. Lastly, class IV bands were lineage independent. These bands can be explained by monophyletic alterations that resulted in loss of a band and by random drift in sequences that were ancestral to both lineages.

In order to identify and quantify each occurrence of bands marking class I to IV events in the OBGS data set, we next devised a method to identify and sort polymorphic addresses in the composite binary data using macro functions in Microsoft Excel to execute the sorting process. The columns of binary data for each strain were first ranked numerically according to the relative position of the strain in the NJ tree in Fig. 1. Polymorphic addresses were then identified, classified, and grouped by class. The image that resulted from sorting polymorphic characters present in at least three strains is shown in Fig. 3.

Like the pairwise alignments of the binary strings from the NJ analysis, the sorting functions identified 258 of the 1,159 total binary addresses that were polymorphic in the O157 strains. Of the 258 polymorphisms, 148 occurred in at least three strains and were further enumerated. Class I characters represented 15 of the 148 polymorphisms. Assuming that the inferred phylogeny is valid, the relatively large number of class I events that marked divergence of the lineages suggests that the lineages are significantly divergent, and the presence of shared class I polymorphisms in both Australian and United States isolates is consistent with the hypothesis that divergence of two lineages was an ancestral event that preceded geographic dissemination. Of the 22 class II polymorphisms, 8 were found exclusively in each member of clade C, D, or E. These clade-specific polymorphisms would be expected if the clades arose through recent clonal expansion of lineage II subpopulations. Since clade D is comprised entirely of O157:H[−] strains from Australia, the geographic, phenotypic, and genotypic similarities suggest that this clade could be a regional subclone unique to Australia. In contrast to the lin-

FIG. 1. Genetic relationships of *E. coli* O157:H7 strains from the United States and Australia. The dendrogram was produced by NJ analysis of binary files representing 1,159 OBGS products from each strain obtained with five OBGS primer combinations (tree length, 1,261; consistency index, 0.6772; retention index, 0.7573; 328 characters). Bootstrap percentages, derived from 1,000 replications of an NJ search, are indicated at the nodes for which the percentages are more than 50%. The tree was rooted by using strain AU133c, a bovine STEC isolate, as the outgroup. The designations of strains originating from Australia and New Zealand begin with AU and NZ, respectively. Human clinical isolates are indicated by boldface type, and their designations end with H.

age II clades, monophyletic markers did not distinguish clades A and B. If OBGS sampling of the genome was random and if the lineage I strains were representative of the populations found on each continent, the limited genetic diversity could indicate that lineage I strains arrived more recently on one or both of the continents studied.

Polyphyletic pattern of genome variation and phylogeographic variation. Relative to the inferred phylogeny, monophyletic markers comprised only about one-fourth of the polymorphisms that were detected and accounted for the phy-

logenetic signal that distinguished only three of the five clades. If we excluded three class III polymorphisms and 11 class IV polymorphisms that could be explained simply by loss of an OBGS product from monophyletic clusters of strains, the remaining 97 polymorphic OBGS products were distributed among polyphyletic groups of strains. The ratio of monophyletic polymorphisms to polyphyletic polymorphisms implies that genome diversification occurred rapidly by random drift and accumulation of common alleles and that selection only periodically fixed alleles in subpopulations.

FIG. 2. Distribution of different classes of polymorphisms. Segments of OBGS images illustrating polymorphic bands of the four distribution classes (classes I to IV) are shown in the four panels. The samples were loaded in phylogenetic order, as listed across the top. Each image segment was reduced fivefold vertically, and the molecular weights for the top and bottom regions of each image segment are indicated on the right. The image segments were obtained by using the following primer combinations: class I, OCT22-OCT6C; class II, OCT22-OCT12C; class III, OCT3B-OCT4C; and class IV, OCT3B-OCT4C. The bands of interest are indicated by arrows.

Since clades A and B are geographically separated and were supported by significant bootstrap values, we next asked whether they could be distinguished by region-specific polymorphisms that were class III or IV polymorphisms. This question was examined by sorting the data first by geography and then by inferred phylogeny. The results obtained when we sorted polymorphic markers that were present in three or more strains and were found only in United States or Australian strains are shown in Fig. 4. We identified 14 polymorphisms that were found only in United States strains and 29 polymorphisms that were found only in Australian strains. Six of the polymorphisms in the United States strains and five of the

polymorphisms in the Australian strains were the same class II polymorphisms shown in Fig. 3 that distinguished the geographically separated clusters of strains in clade C and all of the strains in clade D. The remaining geographic polymorphisms were class III and IV markers; these markers included two class IV markers that were found only in all United States strains and four class IV markers that were present in all clade B strains and most other Australian strains. If it is assumed that the phylogeny is valid, the mono- and polyphyletic patterns of geographic distribution suggest that there are independent modes of phylogeographic variation. The first mode can be explained simply by clonal expansion of a subpopula-

FIG. 3. Sorting of polymorphic OBGS products by phylogeny. Binary coding of the presence or absence of bands in five OBGS reaction mixtures was rendered in phylogenetic order, and the polymorphic positions were grouped by class of distribution by using sorting functions in Microsoft Excel. Polymorphic addresses present in three or more strains were sorted and rendered as cells with different colors depending on the class of distribution pattern. Occurrences of class I polymorphisms are red, occurrences of class II polymorphisms are pink, occurrences of class III polymorphisms are green, and occurrences of class IV polymorphisms are blue. The order of strains is indicated on the left.

FIG. 4. Sorting of polymorphic OBGS products by geography. Binary coding of the presence or absence of bands in five OBGS reaction mixtures was rendered in geographic order and secondarily in phylogenetic order by sorting functions in Microsoft Excel. Only the geography-specific polymorphic addresses present in three or more class strains are shown. The cells containing geographic bands were rendered in different colors by conditional formatting functions according to the class of distribution relative to phylogeny (red, class I; pink, class II; green, class III; blue, class IV). The order of the strains is indicated on the left.

tion, such as clade D, perhaps due to its ability to occupy a unique niche. The second mode appears to arise from concentration of geographic alleles in populations derived from both lineages, probably as a consequence of the populations occupying a regional niche.

DNA sequence analysis of monophyletic lineage markers and clade divergence. In order to identify the nature of the OBGS polymorphisms that resulted in the inferred patterns of

phylogeny, we purified, cloned, and sequenced representative monophyletic polymorphisms that distinguished the two lineages and the three lineage II clades. Bands were purified from OBGS reaction mixtures by using a two-dimensional y-axis scanner to locate the bands in gels in which migration had been stopped prior to migration of the bands past the optical unit of the sequencer. After excision, purification, and reamplification, the labeled reamplified products were then checked for

FIG. 5. Purification and reamplification of the class I OBGS product. The excised class I polymorphic bands obtained with the OCT22–OCT6C primer combination were purified from the reaction products of strains 93-001 and FRIK920 and were reamplified (Reamp) with the same primer combination. The reamplification products were then electrophoresed alongside the original OBGS reaction mixtures to check for purity. The arrows indicate the positions of the reamplified bands that comigrated with the bands in the mother reaction mixture preparations from which they were excised.

purity by electrophoresis alongside the mother reaction mixture preparations from which they were originally excised (Fig. 5). Once their identities were confirmed, the reamplified products were tailed, cloned, and sequenced. The DNA sequences were then compared to the genome sequences of the lineage I EDL933 O157:H7 strain, (Fig. 1), the Sakai O157:H7 strain (17, 27), and the MG1655 K-12 strain (1a).

To identify an event associated with the inferred divergence of the lineages, a pair of class I polymorphisms was chosen from the OCT22–OCT6C combination. The DNA sequences

of bands from strains 93-001 (lineage I) and FRIK920 (lineage II) were nearly identical and spanned the intergenic region between the divergent *folD* and *sfmA* genes. The sequence from 93-001 (accession number AF368044) was identical for the entire length of the fragment to the sequences from the EDL933 and Sakai O157:H7 strains. The lineage II strain sequence (accession number AF368045) had an eight-base duplication and a single additional base flanking the duplication. The duplicated region began 129 bases upstream of *folD* and apparently was a derived region in lineage II since the duplication was not present in the MG1655 sequence. This polymorphism is a convenient marker of the lineage II clone because it has been found in all of the lineage II strains examined but has not been found in the lineage I strains that we have tested to date (Kim and Benson, unpublished data). Whether the duplication actually influences transcription from either of the promoters or confers a significant phenotype requires further study. We designated this marker of lineage divergence 226C-4.

In order to identify events that mark clonal expansion of lineage II clades C, D, and E, we purified and cloned class II polymorphic OBGS products found only in all members of each clade. The clade C-specific band, obtained with primer combination OCT3B–OCT4C, yielded a sequence (accession number AF368051) that showed no significant similarity to sequences in the EDL933 or Sakai genome or to any entries in the BLAST databases. Southern blot analysis using a labeled PCR product amplified from the cloned fragment identified a strongly hybridizing 700-base *EcoRV* fragment that was present in each clade C strain but not in any other strain (data not shown). Evidently, this polymorphism, designated 3B4C-2, marks a gene acquisition event that was associated with divergence of the clade C subpopulation from its lineage II ancestor, or it arose by lateral spread in a subpopulation of lineage II strains.

Two polymorphic bands obtained with the OCT22–OCT6C primer combination were selected as clade D markers. These two bands, approximately 1,100 bases long, had contrasting patterns of distribution. The DNA sequences of the band from clade C strain AU6 (accession number AF368047) and the larger band from clade D strain AU1808 (accession number AF368048) were sequences from the *leuL-leuO* region. The sequence obtained from the clade D band had an additional CTA triplet located in a conserved region consisting of four tandem CTA repeats in the clade C sequence and in the EDL933, Sakai, and MG1655 genome sequences. These triplets comprised leucine codon repeats in the *leuL* leader peptide that are involved in attenuation of *leuABCD* expression (37). Expression of this biosynthetic pathway is modulated, in part, by attenuation, and minor increases in the length of the leucine codon repeats apparently have a modest effect on the threshold level of leucine necessary for induction (1). We designated this derived clade D marker *leuL*(CTA)₅.

DNA sequence analysis of the clade E-specific band, obtained with primer combination OCT3B–OCT4C, showed that it originated from the *gatC* gene, which encodes enzyme IIC of the dulcitol-specific phosphotransferase system transporter. The sequence of the band, designated 3B4C-5 (accession number AF368046), was identical to the EDL933 and Sakai O157:H7 genome sequences for the entire length of the cloned

fragment, including the OCT3B and OCT4C priming sites. Since nucleotides present in the EDL933 and Sakai genome sequences 5' to the OCT3B and OCT4C priming sites do not match the 5' AT spacer dinucleotide in the OBGS primers, it is possible that clade E strains have one or more mutations which bring the adjacent nucleotides into complementarity with the 5' AT dinucleotides of the primers. Alternatively, it is also possible that some epigenetic phenomenon, such as base modification, could prohibit annealing of the primers to non-clade E DNA at these positions.

DNA sequence analysis of polyphyletic genome alterations.

Since clades A and B are distinguished from one another by geographic polymorphisms, we cloned and sequenced a class IV band obtained with primer combination OCT3B-OCT12C, designated 3B12C-2 (accession number AF368050), that was found only in Australian strains except clade D strains AU1808, AU1809, AU1817, and AU1822. There was no apparent band obtained with the same primer combination for the opposing pattern of distribution. The 3B12C-2 sequence aligned with a segment of the EDL933 and Sakai genome sequences spanning the *folB* and *bacA* genes, which encode dihydroneopterin aldolase (16) and a putative undecaprenol kinase that confers resistance to bacitracin at a high copy number, respectively (3). Four single-nucleotide polymorphisms (SNPs) were present in the 3B12C-2 sequence compared to the EDL933 and Sakai sequences, but six SNPs were conserved in the 3B12C-2, EDL933, and Sakai sequences compared to the MG1655 sequence, supporting the notion that unique positions in the 3B12C-2 marker were derived and that the allele was probably not acquired. Two of the derived SNPs in clade B are silent, while the other two result in an isoleucine-to-threonine substitution at position 26 of the *bacA* open reading frame (ORF) and a valine-to-alanine substitution at position 314 of the *folB* ORF. In addition to the internal SNPs, sequences obtained from the 3B12C-3 band and the EDL933 and Sakai genomes also diverged immediately upstream of the OCT3B and OCT12C binding sites, suggesting that mismatches with the AT dinucleotide spacers of the octamer primers are probably the reason why the band did not appear to be amplified from EDL933 and other United States strains or K-12 control strain MC1061 in our OBGS reaction mixtures.

In addition to polyphyletic polymorphisms that distinguished clades A and B, we also sequenced randomly chosen class IV and III bands to gain further insight into the genetic basis for the seemingly large degree of random drift detected by OBGS. One of the class IV OBGS products, obtained with the OCT21-OCT13C combination, yielded a 490-base sequence (accession number AF368052) that had a C-T mismatch at position 321 of the L0079 ORF encoding the *ssb* gene of phage 933W. The mismatch was in the OCT21 octamer priming site and probably prohibited binding of the OCT21 primer in EDL933 and the other 30 strains in our strain set lacking this polymorphic band. However, relative to the Sakai genome, 322 bases at the 3' end of the band aligned with two different prophage elements, but no alignments were observed in the 166 bases at the 5' end, indicating that several different types of events could account for the absence of this band.

The two class III bands that were selected also yielded sequences homologous to prophage and chromosomal genes. A class III polymorphism obtained with the OCT22-OCT12C

combination, present in FRIK920, FRIK1574, FRIK1988, AU514, AU1820, AU1810, and AU1811, yielded a segment from clade C strain FRIK920 (accession number AF368049) that was derived from the *ybgH* gene of the EDL933, Sakai, and MG1655 sequences encoding a putative oligopeptide transport protein. Two SNPs were found in the FRIK920 sequence compared to the O157:H7 genome sequences; one was an A-to-T substitution at position 351 of the *ybgH* coding region, which created a premature stop codon, and the other was a C-to-T substitution 43 bases upstream of the *ybgH* start codon that was within the OCT22 primer itself. In addition to the SNPs found only in FRIK920, 17 additional SNPs were conserved in the FRIK920 sequence and other O157:H7 sequences compared to MG1655, indicating that the FRIK920 allele is likely to be a derived allele. It is tempting to speculate that the random occurrence of the 22-12C marker could reflect a propensity for the -43 C-to-T substitution, perhaps as a result of frequent cytosine deamination, creating an OCT22 priming site.

An additional class III band, 3B4C-1, was obtained with FRIK1988, AU1514, AU1811, and AU1820 in clade E, and this band exhibited a complex pattern of homology to prophages in the EDL933 and Sakai genome sequences. This 311-base fragment, obtained with the OCT3B-OCT4C combination (accession number AY036017), showed significant homology to a 122-base region of the Z1842 ORF of cryptic prophage CP933-C in EDL933 and the corresponding position in the Sakai O157:H7 genome but limited homology in the sequences flanking this segment. An intact OCT4C priming site was present at the corresponding end of the CP933-C sequence but not at the OCT3B priming site. Evidently, the polymorphic band originated from unique prophages residing in the FRIK1988, AU1514, AU1811, and AU1820 genomes or from a recombination event that occurred at some point in the evolution of some clade E strains. The significant diversity of the prophage contents of the O157:H7 genome sequences of EDL933 and Sakai (17, 27) and the diversity that can be determined by restriction fragment length polymorphism analysis with lambdoid phage genome probes (7, 20, 30) suggest that movement and recombination among prophages are significant mechanisms of genome diversification. It will be interesting to conduct a formal analysis of the frequency at which recombination within prophages and insertion-excision of lysogenic bacteriophages participate in genome diversification.

DISCUSSION

Comparative genome analyses of sorbitol-negative, β -glucuronidase-negative *E. coli* O157:H7 strains previously showed that two distinct lineages could be detected in the United States (20). In this study, we found that the divergence of these lineages was probably an ancestral event that preceded geographic dissemination. This conclusion was based on phylogenetic inferences from OBGS data, the distribution of genome alterations detected by OBGS, and the nature of alterations in representative loci that mark the pattern of descent.

Because we sorted markers using the inferred phylogeny and identified conserved polymorphisms using the same phylogeny, the apparent circularity indicates that it is formally possible that some other process is responsible for the distribution of

FIG. 6. Stepwise model of the evolution of O157:H7 and O157:H- EHEC strains. The model is based on that of Feng et al. (11). Each oval represents a progenitor, and evolution proceeds from left to right from ancestral to contemporary states. The A3 to A6 ancestors are designated as described by Feng et al. (11). The markers of events are indicated as follows: Sor⁻, inability to ferment sorbitol; uidA⁻, loss of β-glucuronidase activity; TAI, acquisition of the tellurite resistance-adherence island (35). Specific polymorphisms marking the divergence of A7 and clades A to E are described in the text.

markers. For example, class I and II polymorphisms could be a result of lateral spread of mobile elements among different populations rather than a result of a stepwise pattern of vertical descent. However, the nature of the representative monophyletic alterations that were determined argues against this hypothesis being the sole explanation for the data since the class I polymorphism and two of the three class II polymorphisms are derived states in chromosomal genes and are not linked to mobile elements. Furthermore, sequence analysis of additional class I markers, identified in our laboratory as part of a systematic identification of all conserved markers of lineage divergence, showed that each marker is in an independent locus (Kim and Benson, unpublished), implying that the markers most likely originated through stepwise descent rather than convergent evolution. Although each of the additional class I markers represents an independent event, nearly one-half of them resulted from gene acquisition events, movement of insertion sequences, and movement or recombination within prophage, implying that the conserved polymorphisms comprise a mosaic of mutation, recombination, and acquisition events fixed by one or more selective bottlenecks.

Extended model for evolution of extant O157 EHEC populations. Based on the inferred phylogeny and the evidence that vertical descent of conserved markers occurred, we propose extension of the evolutionary model for contemporary O157:H7 populations (11) to include additional steps subsequent to the emergence of the sorbitol-negative, β-glucuronidase negative ancestor (Fig. 6). The most recent common ancestor of contemporary *E. coli* O157:H7 populations, designated A6 by Feng et al. (11), is believed to have been derived from an LEE⁺ Stx2⁺ O157:H7 ancestor, designated A3, through a stepwise sequence of events that included lysogenization by an *stx1*-converting phage, acquisition of the TAI

locus encoding tellurite resistance and the Iha adhesin, and sequential loss of sorbitol fermentation and β-glucuronidase production (11, 35). Evolution from A3 to A6 can be inferred from multilocus genotyping data, as well as from cumulative character states in the uidA alleles of A3, A5, and A6 descendants (11), culminating in a dinucleotide insertion at position 685 observed in the EDL933 and Sakai genomes that apparently inactivated the gene in A6 descendants (17, 27). Prior to these events, a nonmotile lineage diverged from A3, leading to ancestor A4 of populations of sorbitol-fermenting, β-glucuronidase-positive O157:H- strains that have been found in continental Europe (11, 19).

When our OBGS data are included, an additional population, which descended from the A6 ancestor, must be included to provide a simple explanation for the data. Descendants resulting from this divergence event comprise the lineage I (A6) and lineage II (A7) populations, and assignment of lineage I and II descendants to the A6 (ancestral) and A7 (derived) lineages is based on the DNA sequence of the class I polymorphism, showing that it is a derived polymorphism in lineage II. Consistent with this assignment, three of eight class I polymorphic bands found only in lineage I strains are also obtained with outgroup strain AU133c, and sorting of markers obtained from a larger set of United States strains identified additional class I bands shared by lineage I and K-12 (Wise and Benson unpublished). Given the number of class I OBGS polymorphisms like 226C-4 that mark this event, the A7 ancestor emerged either from a highly divergent subpopulation through a single periodic selection event or descended through sequential bottlenecks. By understanding the nature of class I alterations, we hope to gain further insight into unique physiological and ecological characteristics of A6 and A7 descendants and the events that led to their divergence.

After A6 and A7 populations moved to the different continents, several subpopulations apparently emerged, some of which, such as clades C, D, and E, are marked by monophyletic polymorphisms detected by OBGS. Since multiple class II markers were detected in each of these clades (Fig. 3) and since no class II markers were found only in two of the three clades, we suggest that each clade may have arisen independently from the common A7 ancestor, perhaps through independent selective bottlenecks. Clade D may have arisen through a regional bottleneck since it comprises only Australian strains. This clade is distinct from the populations of non-motile O157 EHEC that descended from A4 because of its sorbitol and β -glucuronidase phenotypes and the presence of the TAI island, which was detected in all descendants of A6 and A7 in this strain set (Nietfeldt and Benson, unpublished data). Examining the distribution of the clade D markers, such as *leuL*(CTA)₅, in non-sorbitol-fermenting, β -glucuronidase-negative O157 EHEC strains should provide data regarding the geographic distribution of this clade.

In contrast to the A7 descendants, the A6 descendants in clades A and B showed only limited genetic diversity and were differentiated from each other primarily by sets of polymorphisms that were polyphyletic but conserved or nearly conserved with respect to geography. We speculate that this pattern of diversity reflects regional selection imposed on populations of lineage I strains that recently arrived on either or both of the continents studied. Testing this hypothesis, however, will require analysis of larger sets of strains.

Distribution of human and bovine isolates. Although descendants of both the A6 (lineage I) and A7 (lineage II) ancestors were detected among the Australian strains, most of the strains tested, including several human clinical clade D and E strains from Australia, were derived from the A7 ancestor. This finding is in contrast to the results of OBGS analysis of human and bovine O157:H7 strains from the United States, where few human isolates were observed among lineage II strains (20). There are several possible explanations for the observed differences. First, although in these studies we used strains that are temporally and spatially diverse in terms of their origins, it is possible that the differences reflect bias in the relatively small sets of strains. A second possibility is that differences in animal production and food preparation practices between the two countries studied contribute to bias in transmission. And finally, it is also possible that clade D and E strains have virulence properties that are distinct from those of their clade C counterparts. It is worth noting that most bovine strains from the United States that have been tested by the OBGS method cluster with the clade C strains, while the United States clade E strain is closely related to the few human lineage II isolates identified in the United States studies (20). We are currently devising a multiplex assay, based on class I and II polymorphisms, which will provide a rapid method for strain classification that can be used to statistically evaluate the differential virulence hypothesis.

Generation of genome diversity. Although the small set of class III and IV polymorphisms that we sequenced does not allow quantitative conclusions, the fact that we identified alterations in two different genomic sequences and two different prophage sequences implies that variation in each sequence contributed to genome diversification. The polyphyletic alter-

ations detected in genomic sequences appear to be SNPs that create or destroy OBGS primer binding sites, and the results of comparisons of SNPs inside these regions are consistent with the hypothesis that the alleles are derived. Among the polyphyletic alterations detected in prophage sequences, we found one example of a polymorphism that arose either from lysogenization by a unique phage or by a recombination event that generated a chimeric segment in a resident cryptic prophage. Due to the high degrees of similarity among prophages and cryptic prophages that reside in O157:H7 genomes (17, 27), the diversity of prophage contents of different strains (7, 17, 20, 27, 30), and the modular nature of phage genomes (2, 4, 5), it is difficult to determine whether polymorphic OBGS products carrying chimeric prophage sequences arise from integration and excision of lysogenic phages or by recombination. Several of the class III and IV polymorphisms can be found in common sets of polyphyletic strains, which supports the view that the polymorphisms are physically linked and arise from movement of phages into and out of the genome. On the other hand, recombination events in prophage sequences could also appear as linked sets of class III and IV polymorphisms since highly conserved modules of phage genomes from similar and diverse hosts can vary significantly in length (2, 4, 5, 12, 18). Regardless of the precise mechanism of phage-mediated variation, it seems that infection of strains by bacteriophages is common and offers significant opportunities for lysogenization and recombination events to occur since the populations occupy phage-dense environments, such as the rumen.

Whether diversification occurs in prophage or genomic sequences, some of the drift may not be present in entirely random groups of strains. As demonstrated by sorting of the OBGS polymorphisms primarily by geography (Fig. 4), we identified different types of phylogeographic variation. The apparent clonal expansion of clade D, which comprises only Australian strains, is the first type. In contrast, the geography-specific class IV polymorphisms that distinguish clades A and B may be a second mechanism of phylogeographic variation, in which provincial selective pressures concentrate regional alleles or common alterations in genetically distinct populations that occupy a common niche. Such information is not readily apparent when data are sorted by inferred phylogeny alone. The fact that geographic distributions of class IV markers can be detected brings up the possibility that additional patterns of selection and/or allele concentration may be represented among the class III and IV alterations and may be previously unrecognized patterns of selection or common ecological characteristics. Implementing additional algorithms for data sorting and analysis of extensive OBGS fragment data may provide additional tools to identify shared physiological and ecological characteristics of strain sets and may ultimately help workers identify unique signatures of selection.

ACKNOWLEDGMENTS

We thank Lawrence Harshman and Robert Hutkins for insightful discussions and critical reviews of the manuscript. We are indebted to Roy Robins-Browne and Dianne Lightfoot for sharing human clinical isolates from their collections.

This research was funded by grant 98-35291-6215 from the United States Department of Agriculture National Research Initiative Com-

petitive Grants Program to A.K.B. and by Nebraska Legislative Bill LB1206.

REFERENCES

- Bartkus, J. M., B. Tyler, and J. M. Calvo. 1991. Transcription attenuation-mediated control of *leu* operon expression: influence of the number of Leu control codons. *J. Bacteriol.* **173**:1634–1641.
- Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453–1462.
- Brussow, H., A. Bruttin, F. Desiere, S. Lucchini, and S. Foley. 1998. Molecular ecology and evolution of *Streptococcus thermophilus* bacteriophages—review. *Virus Genes* **16**:95–109.
- Cain, B. D., P. J. Norton, W. Eubanks, H. S. Nick, and C. M. Allen. 1993. Amplification of the *bacA* gene confers bacitracin resistance to *Escherichia coli*. *J. Bacteriol.* **175**:3784–3789.
- Campbell, A. M. 1994. Comparative molecular biology of lambdoid phages. *Annu. Rev. Microbiol.* **48**:193–222.
- Campbell, A. M. 1988. Phage evolution and speciation, p. 1–14. In R. Calendar (ed.), *The bacteriophages*. Plenum Press, New York, N.Y.
- Caprioli, A., and A. E. Tozzi. 1998. STEC infections in continental Europe, p. 38–48. In J. B. Kaper and A. D. O'Brien (ed.), *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains. American Society for Microbiology, Washington, D.C.
- Datz, C. M., C. Janetzki-Milkman, S. Franke, F. Gunzer, H. Schmidt, and H. Karch. 1996. Analysis of the enterohemorrhagic *Escherichia coli* O157 DNA region containing lambdoid phage gene p and Shiga-like toxin structural genes. *Appl. Environ. Microbiol.* **62**:791–797.
- Donnenberg, M. S., S. Tzipori, M. L. McKee, A. D. O'Brien, J. Alroy, and J. B. Kaper. 1993. The role of the *eae* gene of enterohemorrhagic *Escherichia coli* in intimate attachment in vitro and in a porcine model. *J. Clin. Invest.* **92**:1418–1424.
- Dytoc, M., R. Soni, F. Cockerill III, J. De Azavedo, M. L. Louie, J. Brunton, and P. Sherman. 1993. Multiple determinants of verotoxin-producing *Escherichia coli* O157:H7 attachment-effacement. *Infect. Immun.* **61**:3382–3391.
- Faith, N. G., J. A. Shere, R. Brosch, K. W. Arnold, S. E. Ansay, M.-S. Lee, J. B. Luchansky, and C. W. Kaspar. 1996. Prevalence and clonal nature of *Escherichia coli* O157:H7 on dairy farms in Wisconsin. *Appl. Environ. Microbiol.* **62**:1519–1525.
- Feng, P., K. A. Lampel, H. Karch, and T. S. Whittam. 1998. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J. Infect. Dis.* **177**:1750–1753.
- Ford, M. E., G. J. Sarkis, A. E. Belanger, R. W. Hendrix, and G. F. Hatfull. 1998. Genome structure of mycobacteriophage D29: implications for phage evolution. *J. Mol. Biol.* **279**:143–164.
- Gouveia, S., M. E. Proctor, M.-S. Lee, J. B. Luchansky, and C. W. Kaspar. 1998. Genomic comparisons and Shiga toxin production among *Escherichia coli* O157:H7 isolates from a day care center outbreak and sporadic cases in southeastern Wisconsin. *J. Clin. Microbiol.* **36**:727–733.
- Griffin, P. M., and R. V. Tauxe. 1991. The epidemiology of infections caused by *Escherichia coli* O157:H7, other enterohemorrhagic *E. coli*, and the associated hemolytic uremic syndrome. *Epidemiol. Rev.* **13**:60–98.
- Gunzer, F., H. Bohm, H. Russmann, M. Bitzan, S. Aleksic, and H. Karch. 1992. Molecular detection of sorbitol-fermenting *Escherichia coli* O157 in patients with hemolytic-uremic syndrome. *J. Clin. Microbiol.* **30**:1807–1810.
- Haubmann, C., F. Rodich, E. Schmidt, A. Bacher, and G. Richter. 1998. Biosynthesis of pteridines in *Escherichia coli*. *J. Biol. Chem.* **273**:17418–17424.
- Hayashi, T., K. Makino, M. Onishi, K. Kurokawa, K. Ishi, K. Yokoyama, C.-G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasaki, N. Ogasawara, T. Yasunaga, S. Kuhara, T. Shiba, M. Hattori, and H. Shinagawa. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain. *DNA Res.* **8**:11–22.
- Hendrix, R. W., M. C. M. Smith, R. N. Burns, M. E. Ford, and G. F. Hatfull. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc. Natl. Acad. Sci. USA* **96**:2192–2197.
- Karch, H., H. Böhm, H. Schmidt, F. Gunzer, S. Aleksic, and J. Heesemann. 1993. Clonal structure and pathogenicity of Shiga-like toxin-producing, sorbitol-fermenting *Escherichia coli* O157:H-. *J. Clin. Microbiol.* **31**:1200–1205.
- Kim, J., J. Nietfeldt, and A. K. Benson. 1999. Octamer based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle. *Proc. Natl. Acad. Sci. USA* **96**:13288–13293.
- Lee, M.-S., C. W. Kaspar, R. Brosch, J. Shere, and J. B. Luchansky. 1996. Genomic analysis using pulsed-field gel electrophoresis of *Escherichia coli* O157:H7 isolated from dairy calves during the United States national dairy heifer evaluation project (1991–1992). *Vet. Microbiol.* **48**:223–230.
- Levine, M. M. 1987. *Escherichia coli* that cause diarrhea: enterotoxigenic, enteropathogenic, enteroinvasive, enterohemorrhagic, and enteroadherent. *J. Infect. Dis.* **155**:377–389.
- McDaniel, T. K., K. G. Jarvis, M. S. Donnenberg, and J. B. Kaper. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proc. Natl. Acad. Sci. USA* **92**:1664–1668.
- McGraw, E. A., J. Li, R. K. Selander, and T. S. Whittam. 1999. Molecular evolution and mosaic structure of alpha, beta, and gamma intimins of pathogenic *Escherichia coli*. *Mol. Biol. Evol.* **16**:12–22.
- Newland, J. W., N. A. Stockbrine, F. F. Miller, A. D. O'Brien, and R. K. Holmes. 1985. Cloning of shiga-like toxin genes from a toxin converting phage of *Escherichia coli*. *Science* **230**:179–181.
- O'Brien, A. D., L. R. Marques, C. F. Kerry, J. W. Newland, and R. K. Holmes. 1989. Shiga-like toxin converting phage of enterohemorrhagic *Escherichia coli* strain 933. *Microb. Pathog.* **6**:381–390.
- Perna, N. T., G. Plunkett III, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, J. Gregor, H. A. Kirkpatrick, G. Posfai, J. Hackett, S. Klink, A. Boutin, Y. Shao, L. Miller, E. J. Grotbeck, N. W. Davis, A. Lim, E. T. Dimalanta, K. D. Potamou, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner. 2001. Genome sequence of enterohemorrhagic *Escherichia coli* O157:H7. *Nature* **409**:529–533.
- Reid, S. D., C. J. Herbelin, A. C. Brumbaugh, R. K. Selander, and T. S. Whittam. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**:64–67.
- Robins-Browne, R., E. Elliott, and P. Desmarchelier. 1998. Shiga toxin-producing *Escherichia coli* in Australia, p. 66–72. In J. B. Kaper and A. D. O'Brien (ed.), *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains. American Society for Microbiology, Washington, D.C.
- Samadpour, M., L. M. Grimm, B. Desai, D. Alfi, J. E. Ongerth, and P. I. Tarr. 1993. Molecular epidemiology of *Escherichia coli* O157:H7 strains by bacteriophage lambda restriction fragment length polymorphism analysis: application to a multistate foodborne outbreak and a day-care center cluster. *J. Clin. Microbiol.* **31**:3179–3183.
- Schmidt, H., L. Beutin, and H. Karch. 1995. Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL933. *Infect. Immun.* **63**:1055–1061.
- Schmidt, H., B. Henkel, and H. Karch. 1997. A gene cluster related to type II secretion pathway operons of gram-negative bacteria is located on the large plasmid of enterohemorrhagic *Escherichia coli* O157 strains. *FEMS Microbiol. Lett.* **148**:265–272.
- Shere, J. A., K. J. Bartlett, and C. W. Kaspar. 1998. Longitudinal study of *Escherichia coli* O157:H7 dissemination on four dairy farms in Wisconsin. *Appl. Environ. Microbiol.* **64**:1390–1399.
- Swofford, D. L. 1998. PAUP, version 4.0, beta 8. Sinauer and Associates, Sunderland, Mass.
- Tarr, P. I., S. S. Bilge, J. C. Vary, Jr., S. Jelacic, R. C. Habeeb, T. R. Ward, M. R. Baylor, and T. E. Besser. 2000. Iha: a novel *Escherichia coli* O157:H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure. *Infect. Immun.* **68**:1400–1407.
- Tarr, P. I., L. M. Schoening, Y.-L. Yea, T. R. Ward, S. Jelacic, and T. S. Whittam. 2000. Acquisition of the *rfb-gnd* cluster in evolution of *Escherichia coli* O55 and O157. *J. Bacteriol.* **182**:6183–6186.
- Wessler, S. R., and J. M. Calvo. 1981. Control of *leu* operon expression in *Escherichia coli* by a transcription attenuation mechanism. *J. Mol. Biol.* **149**:579–597.
- Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Orskov, I. Orskov, and R. A. Wilson. 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. *Infect. Immun.* **61**:1619–1629.
- Whittam, T. S. 1998. Evolution of STEC strains, p. 195–209. In J. B. Kaper and A. D. O'Brien (ed.), *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains. American Society for Microbiology, Washington, D.C.