Evolution of Genomic Content in the Stepwise Emergence of Escherichia coli O157:H7

Article *in* Journal of Bacteriology · March 2005

DOI: 10.1128/JB.187.5.1783-1791.2005 · Source: PubMed

4 authors, including:

Lukas Wick
ETH Zurich
19 PUBLICATIONS **574** CITATIONS

SEE PROFILE

JOURNAL OF BACTERIOLOGY, Mar. 2005, p. 1783–1791 Vol. 187, No. 5
0021-9193/05/$08.00+0 doi:10.1128/JB.187.5.1783–1791.2005

Evolution of Genomic Content in the Stepwise Emergence of *Escherichia coli* O157:H7†

Lukas M. Wick, Weihong Qi, David W. Lacher, and Thomas S. Whittam*

Microbial Evolution Laboratory, National Food Safety and Toxicology Center, Michigan State University, East Lansing, Michigan

Received 4 October 2004/Accepted 30 November 2004

Genome comparisons have demonstrated that dramatic genetic change often underlies the emergence of new bacterial pathogens. Evolutionary analysis of *Escherichia coli* O157:H7, a pathogen that has emerged as a worldwide public health threat in the past two decades, has posited that this toxin-producing pathogen evolved in a series of steps from O55:H7, a recent ancestor of a nontoxigenic pathogenic clone associated with infantile diarrhea. We used comparative genomic hybridization with 50-mer oligonucleotide microarrays containing probes from both pathogenic and nonpathogenic genomes to infer when genes were acquired and lost. Many ancillary virulence genes identified in the O157 genome were already present in an O55:H7-like progenitor, with 27 of 33 genomic islands of >5 kb and specific for O157:H7 (O islands) that were acquired intact before the split from this immediate ancestor. Most (85%) of variably absent or present genes are part of prophages or phage-like elements. Divergence in gene content among these closely related strains was ~140 times greater than divergence at the nucleotide sequence level. A >100-kb region around the O-antigen gene cluster contained highly divergent sequences and also appears to be duplicated in its entirety in one lineage, suggesting that the whole region was cotransferred in the antigenic shift from O55 to O157. The β-glucuronidase-positive O157 variants, although phylogenetically closest to the Sakai strain, were divergent for multiple adherence factors. These observations suggest that, in addition to gains and losses of phage elements, O157:H7 genomes are rapidly diverging and radiating into new niches as the pathogen disseminates.

Enterohemorrhagic *Escherichia coli* (EHEC) strains were first identified as etiological agents of bloody diarrhea and hemolytic uremic syndrome in the early 1980s and have since been recognized worldwide as a cause of food- and waterborne infectious diseases (9, 14, 16, 22). Two factors critical in the full virulence of EHEC O157:H7 are the locus of enterocyte effacement (LEE), a pathogenicity island that mediates the intimate attachment of bacterial cells to the intestinal epithelium, and the production of one or more Shiga toxins (4, 14, 26). Besides these two well-studied virulence factors and the large virulence plasmid (pO157), many other putative virulence genes have been identified since the completion of genome sequences of two O157:H7 strains (11, 31).

Evolutionary analysis has shown that O157:H7 strains are genetically most closely related to enteropathogenic *E. coli* O55:H7 strains (40) and has engendered a model (7) specifying that O157:H7 evolved through a series of transitional steps from a nontoxigenic progenitor (Fig. 1A). The model predicts that the most recent common ancestor (A1 in Fig. 1) of today's O157:H7 and O55:H7 strains contained the LEE and presumably could elicit diarrhea via an attachment-effacement mechanism. In addition, the ancestor resembled wild-type *E. coli* in its abilities to ferment sorbitol (SOR^+) and express β-glucuronidase (GUD^+). In a first step towards O157:H7, A1 acquired Stx2, presumably through transduction, resulting in a Stx2-positive O55:H7 (A2 in Fig. 1). In the next step, the large virulence plasmid (pO157) was gained and the somatic antigen switched from O55 to O157. From this stage (A3), two separate lines evolved. One branch lost motility by mutation in the flagellar operon (23), resulting in the SOR^+ O157 (also called SF O157) clone discovered in hemolytic uremic syndrome cases in Germany (15, 17) and hereafter referred to as the German clone (A4). The other branch, from which the GUD^+ O157 strains descended, lost the ability to ferment sorbitol and gained Stx1 (A5). Subsequently, mutational inactivation of the *uidA* gene (24) resulted in the non-sorbitol-fermenting, β-glucuronidase-negative phenotype typical of *E. coli* O157:H7 (A6). It was the clonal descendants of A6 that expanded and spread geographically and that now account for most disease caused by EHEC (19).

Comparison of genomic sequences from two pathogenic O157:H7 strains (11, 31) and avirulent *E. coli* K-12 (2) uncovered hundreds of genes that are present only in the pathogens (O islands) or only in K-12 (K islands). The objective of this study was to elucidate at what points in the evolution of O157:H7 were O islands, especially the ones encoding putative virulence factors, gained and K islands lost. To this end, the total gene contents of strains representing different stages in the stepwise evolution model were assessed by comparative genomic hybridization analysis with spotted-oligonucleotide glass arrays. The arrays (*E. coli* O157 arrays; MWG Biotech, High Point, N.C.) are multigenome arrays that contain 50-mer probes targeting a combined total of 6,176 open reading frames (ORFs) from the *E. coli* K-12 (2) and two O157:H7 genomes (11).

* Corresponding author. Mailing address: Microbial Evolution Laboratory, 165 Food Safety & Toxicology Building, Michigan State University, East Lansing, MI 48824. Phone: (517) 432-3100, ext 178. Fax: (517) 432-2310. E-mail: whittam@msu.edu.

† Supplemental material for this article may be found at http://jb.asm.org/.

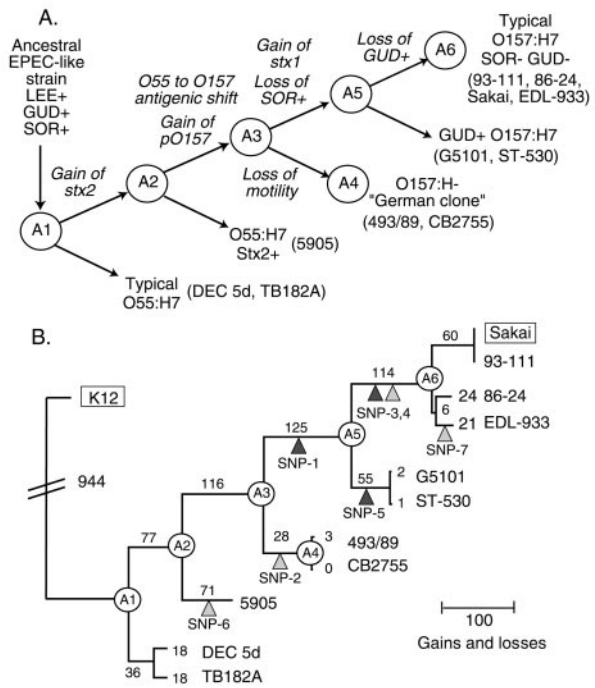

FIG. 1. Evolutionary genomic changes in the emergence of *E. coli* O157:H7. (A) Stepwise model for the evolution of *E. coli* O157:H7 from an enteropathogenic *E. coli*-like ancestor (modified from reference 7 with permission of the publisher). (B) Maximum-parsimony tree (21) based on the presence and absence of genes inferred by microarray hybridizations for 5,121 genes. Boxes indicate the reference genomes. The number of events for each branch is based on the assumption that gene gains and losses occur independently. Triangles mark occurrences of seven SNPs (gray, synonymous; black, nonsynonymous) found by sequencing 7,470 bp in 15 conserved genes. Note that the inferred topology based on gene content is identical to the model in panel A.

MATERIALS AND METHODS

Strains. The pathogenic *E. coli* strains examined in this study are listed in Table 1. Genomic DNAs of the sequenced strains O157:H7 RIMD 0509952 (Sakai) (11) and nonpathogenic *E. coli* MG1655 (K-12) (2) were used as references in two-color hybridization experiments.

Oligonucleotide arrays. We used MWG *E. coli* O157 arrays (MWG Biotech), which contain 6,176 50-mer oligonucleotides specific for three *E. coli* strains, K-12 (MG1655), Sakai, and EDL-933. The array also contains 45 oligonucleotides in replicate and 68 *Arabidopsis* control oligonucleotides. To verify the probes with the up-to-date genome annotations, we compared by BLAST analysis all 6,176 probe sequences of the MWG array against the three *E. coli* genomes (K-12, Sakai, and EDL-933) and recorded the two highest hits for every probe (top hit and second hit) for each genome. A probe was considered specific for a target when its top hit had 40 bp or more that were identical to the 50-bp sequence stretch in the strain (≥80%). We excluded probes that showed potential for nonspecific hybridizations (hits of 37 to 39 bp of overall identity or hits of 25 to 36 bp of overall identity including a stretch of 15 or more consecutive base pairs with 100% identity) or multiple target hybridizations with K-12 or Sakai DNA (because these two strains were used as reference genomes in this study). With respect to the K-12 and Sakai genomes, out of the 6,176 probes, 14 had no target (EDL-933 specific), 38 had a potential for nonspecific hybridization, 433 had multiple targets (with 365 of these belonging to phage or phage-like elements), and 5,691 matched single genome targets. Of these 5,691 probes, 3,963 target both genomes, 1,257 target only Sakai, and 471 target only K-12. All probes were assigned ORF designations (b-, ecs-, or z- numbers) or intergenic region labels based on the RefSeq database available on the National Center for Biotechnology Information website (38). Of these 5,691 probes, 5,353 target the same genomes (Sakai, K-12, or both) as given in the original annotation by MWG Biotech, and 338 show differences.

TABLE 1. Pathogenic *E. coli* strains used in this study

Strain	Serotype	Origin, yr	Reference
DEC 5d	O55:H7	Sri Lanka, 1965	32
TB182A	O55:H7	United States (Washington), 1991	2a
5905	O55:H7	United States, 1994	32
493/89	O157:H-	Germany, 1989	32
CB2755	O157:H-	Germany, 1993	33
G5101	O157:H7	United States (Washington), 1995	11a
ST-530	O157:H7	United States (Michigan), 2003	This study
93-111	O157:H7	United States (Washington), 1993	32
86-24	O157:H7	United States (Washington), 1986	35a
Sakai	O157:H7	Japan (Sakai), 1996	11

DNA labeling. Genomic DNA was sheared into 500- to 5,000-bp fragments in a cup sonicator (Heat Systems Ultrasonics W-225; 20 kHz, 200 W). A total of 250 ng of sheared DNA was aminoallyl-dUTP (Sigma, St. Louis, Mo.) labeled with the Invitrogen (Carlsbad, Calif.) DNA labeling system, using a modified 25× deoxynucleoside triphosphate mix consisting of 12.5 mM (each) dATP, dGTP, and dCTP; 2.1 mM dTTP; and 10.4 mM aminoallyl-dUTP. The DNA was purified with Qiagen (Valencia, Calif.) PCR purification columns, using modified amine-free wash (5 mM KPO_4, pH 8.0, 80% ethanol) and elution (5 mM KPO_4 [pH 8.0]) buffers. The aminoallyl-labeled DNA was dried down in a vacuum centrifuge, suspended in 4.5 μl of 0.1 M Na_2CO_3 (pH 9.3), Cy coupled by the addition of 4.5 μl of Cy3 or Cy5 dye in dimethyl sulfoxide (1/16 of one vial of Mono-reactive Cy dye [Amersham, Piscataway, N.J.]), and incubated for 1 h at room temperature in the dark. Cy-labeled DNA was purified with Qiagen PCR purification columns. DNA and dye concentrations were determined with a spectrophotometer (NanoDrop Technologies, Rockland, Del.), and labeled DNA was dried down by vacuum centrifugation.

Microarray hybridizations and data processing. Equal amounts of DNAs from strains to be compared, labeled with different Cy dyes, were suspended and combined in a final volume of 35 μl of formamide-based hybridization buffer (MWG Biotech). MWG *E. coli* O157 arrays were hybridized and washed according to the manufacturer's instructions for hybridization with coverslips. Lifter slips (22x40I-2-4710) from Erie Scientific Company (Portsmouth, N.H.) were used. The used arrays were stripped and reused once. For stripping, the arrays were washed two or three times for 5 min each in 90°C H_2O and twice for 10 s each at room temperature in H_2O and then were dried by centrifugation (3 min at 500 × *g*). Test strains were hybridized twice with Sakai as a reference: once on a new chip with the test strain Cy3 labeled and Sakai Cy5 labeled and once on a used chip with the test strain Cy5 labeled and Sakai Cy3 labeled.

Arrays were scanned with a Genepix 4000B instrument (Axon Instruments, Union City, Calif.), and probe intensities (median pixel intensities) were retrieved with Genepix 3.0 software (Axon Instruments). The data quality and the normalization effects were assessed by viewing plots of M versus A [$M = \log_2$ (test/reference); $A = \log_2$ (test × reference)/2] and by checking for spatial effects with GeneTraffic (Iobion, La Jolla, Calif.) and MAANOVA (41) software. Arrays were generally normalized by global LOWESS normalization, unless they showed spatial bias (in which case subgrid LOWESS normalization was used) or unless normalization skewed the data in the plot of M versus A (in which case raw values were used).

Data analysis. Data points were filtered for further analysis if probes showed either printing abnormalities or exhibited a low signal in hybridization with the Sakai reference strain. After filtering and normalization, hybridization data were analyzed as the distribution of the two-color signal ratios by using GACK (Genomotyping Analysis by Charlie Kim) (18). The GACK program uses the shape of the $\log_2$ distributions to locate signal ratio cutoffs for classifying genes as present in a genome or absent. For each array, analyses of the $\log_2$ (test strain/reference strain) distribution ($GACK_1$) as well as of the reciprocal ratio, $\log_2$ (reference strain/test strain) ($GACK_2$), were performed. The $GACK_2$ value provides information about probes without targets in the reference strain and might also detect duplicated targets. We classified genes with a $GACK_1$ value of <-0.4 as absent and those with a $GACK_1$ value of ≥ -0.4 as present. Genes with a $GACK_2$ value of <-0.4 were classified as duplications (see also "Chip validation" below). For probes without targets in the reference strain, $GACK_2$ values of <-0.4 indicate presence in the test strain if the signal exceeds the low-intensity cutoff. Otherwise these genes were also classified as absent in the test strain.

For probes without targets in the reference strain, an additional indirect analysis was done. The $\log_2$ (test strain/K-12) values were calculated from the test

strain-Sakai and the K-12–Sakai hybridizations as follows: $\log_2$ (test/K-12) = $\log_2$ (test/Sakai) − $\log_2$ (K-12/Sakai). GACK analysis was done with these $\log_2$ (test/K-12) values.

Multilocus sequence analysis. The nucleotide sequences of internal fragments of multiple housekeeping genes were determined as described previously (12, 32). The 15 loci were *arcA*, *aroE*, *aspC*, *clpX*, *cyaA*, *dnaG*, *fadD*, *grpE*, *icdA*, *lysP*, *mdh*, *mtlD*, *mutS*, *rpoS*, and *uidA*. The sequencing protocols are available on the STEC website (http://www.shigatox.net/mlst).

Phylogenetic analysis. A phylogeny for the genomes was inferred by parsimony analysis of the presence or absence of genes by using the PAUP (34) program (version 4.0b10). The presence or absence of individual genes was coded as 0 (absence) or 1 (presence) in binary characters. Parsimony analysis was based on the subset of genes that were phylogenetically informative, using the ordinary parsimony algorithm with all steps counted and random sequence addition. Character states were unordered and given equal weight. All genes that were variably absent or present were included in a second parsimony analysis with MEGA (21) to infer the total number of gene gains and losses in genome divergence.

RESULTS

Chip validation. To gauge the accuracy of the oligoarray in assessing genomic content, we validated the chip by comparative hybridization with genomic DNAs from Sakai and K-12. BLAST analysis predicts that 5,691 probes have exactly one target in either genome and that 3,963 (69.8%) occur in both genomes. Of these, 3,693 (93.4%) are identical in sequence in both genomes, whereas 145 probes are more similar in sequence to Sakai than to K-12 targets (Sakai-like probes) and 125 are more similar to K-12 than to Sakai targets (K-12-like probes). In addition, there are 1,728 single-target, strain-specific probes, 1,257 of which occur in Sakai only and 471 of which occur in K-12 only.

We compared the in silico analysis described above with the performance of actual two-color hybridizations with Cy5-labeled Sakai DNA and Cy3-labeled K-12 DNA. The log intensity ratio (*M*) plotted against the mean log intensity (*A*) shows outstanding separation of the strain-specific probes (Fig. 2). We adjusted the cutoff values in separate GACK analyses, with Sakai [$\log_2$ (K-12/Sakai)] and K-12 [$\log_2$ (Sakai/K-12)] as references, to classify genes as present or absent. This analysis used only probes with identical targets in both strains or with targets in only one strain (Fig. 2). A GACK cutoff of −0.4 (Fig. 2) gave <0.6% false negatives (i.e., genes known to be present but classified by the GACK value as absent) and a maximum sensitivity and specificity (1) in distinguishing genes that are present from those that are absent (Table 2). Targets with sequence differences between the genomes are more difficult to classify, because their signals overlap with both present and absent targets (Fig. 2). The sequence similarity within these 269 divergent 50-mer probes ranges from 84 to 98%. Most targets with 96 to 98% similarity are called present, and we estimate that about 50% of targets with 94% similarity are called divergent at a GACK cutoff of −0.4 (see Fig. S6 in the supplemental material).

For probes known to target a reference strain, $GACK_2$ values of <0.5 indicate the possibility that targets have been duplicated or exist in multiple copies in the test strain. In the MWG array probe set, there are 17 probes with multiple targets in Sakai compared to K-12 and seven with multiple targets in K-12 compared to Sakai. (Multiple targets are defined as targets present exactly once in one genome and at least twice, with not more than a 1-bp difference, in the other genome.) If a $GACK_2$ value of less than our cutoff of −0.4 is scored as a multiple hit, 1 of the 17 Sakai multiple targets (6% false negatives) and 1 of the 7 K-12 multiple targets (14% false negatives) are undetected. If only duplications are considered, 1 of 11 duplications in Sakai (9% false negatives) and the single duplication in K-12 (100% false negatives) are not detected. Because the sample size is small, this estimation of false negatives is inexact, but it indicates that most duplications should be detected. The false-positive rate of duplications can be

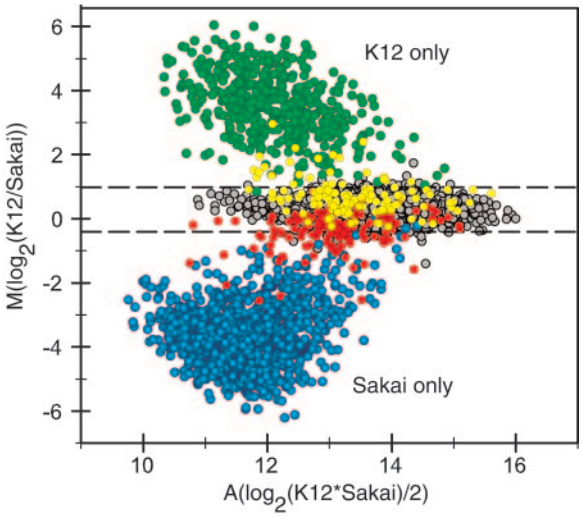

FIG. 2. Plot of *M* versus *A* for two-color hybridization of Sakai and K-12 genomic DNAs for 5,667 probes (24 of the 5,691 single-copy probes were excluded because of poor signals). Probes are placed into five groups (color coded) based on in silico analysis of single-copy targets in the Sakai and K-12 genomes. Gray, identical target sequences in both genomes (n = 3,684); blue, Sakai-only targets (n = 1,246); green, K-12-only targets (n = 468); red, Sakai-like targets (n = 145); yellow, K-12-like targets (n = 124). In the last two groups, homologous target sequences occur in both genomes but have diverged in sequence from 1 to 8 bp. Lines show the cutoffs (GACK value of −0.4) used for calls of present or absent. Targets between the lines are classified by hybridization as present in both strains; others are classified as present in one strain only.

TABLE 2. Sensitivity and specificity analysis for different GACK cutoff values

Reference genome	GACK cutoff value	No.				Sensitivity	Specificity
		Positive		Negative			
		True	False	True	False		
Sakai	−0.5	3,678	5	1,241	6	0.998	0.996
	−0.4	3,674	3	1,243	10	0.997	0.998
	−0.2	3,654	3	1,243	30	0.992	0.998
	0	3,634	2	1,244	50	0.986	0.998
	0.2	3,612	1	1,245	72	0.98	0.999
	0.4	3,502	1	1,245	182	0.951	0.999
	0.5	3,413	1	1,245	271	0.926	0.999
K-12	−0.5	3,680	4	464	4	0.999	0.991
	−0.4	3,662	1	467	22	0.994	0.998
	−0.2	3,631	1	467	53	0.986	0.998
	0	3,594	0	468	90	0.976	1
	0.2	3,532	0	468	152	0.959	1
	0.4	3,444	0	468	240	0.935	1
	0.5	3,391	0	468	293	0.92	1

TABLE 3. Numbers of backbone and phage-related genes that are conserved (present in all) or variably absent or present among recent relatives of *E. coli* O157:H7

Function (*n*)	No.		% VAP
	Conserved (*n* = 4,230)	VAP (*n* = 580)	
Backbone (3,669)	3643	26	0.7
Sakai genes (1,084)	425	34	7.4
Sakai phage-related genes	162	463	74.1
K-12 genes (368)[a]	215	28	11.5
K-12 phage-related genes	96	29	23.2

[a] Conserved K-12 genes absent in all O157:H7 relatives; VAP K-12 genes are present in some strains

estimated with more accuracy; it is the same as the false-negative rate (1 − sensitivity) given in Table 2, which at a cutoff of −0.4 is only around 0.5% (e.g., a gene present in K-12 having a GACK value of <-0.4 with Sakai as a reference strain is called absent in K-12 [false negative], but the same signal is interpreted as a duplication in Sakai when K-12 is used as a reference strain).

Gene content in the stepwise evolution of O157:H7. Genomic DNAs from nine strains representing different stages in the stepwise evolution model of O157:H7 (Fig. 1A) were labeled and hybridized to the oligoarray. All strains were tested twice with a dye swap, using Sakai genomic DNA for a reference. GACK values were calculated to classify genes as present or absent (cutoff = −0.4). From the nine experiments, the normalized $\log_2$ signal ratios (mean ± standard deviation [SD]) corresponding to the GACK value of −0.4 were −0.5 ± 0.04 ($GACK_1$) and −0.48 ± 0.08 ($GACK_2$) for the new arrays and −0.61 ± 0.12 (for both $GACK_1$ and $GACK_2$) for the used arrays. The reproducibilities for the nine dye-swap experiments were 99.0% ± 0.6% (mean ± SD) with binary classification (0 = absent; 1 = present) and 98.1% ± 1.1% (mean ± SD) with trinary classification (0 = absent; 1 = present; 2 = duplicated). Out of 5,691 probes, 5,121 (90%) gave perfectly consistent results, with identical binary classification in the replicate experiments throughout all experiments. This set of high-quality data for all strains was used in the subsequent analysis. Final data sets, including trinary scores, are available in Table S5 in the supplemental material.

A total of 4,230 (82.6%) of the 5,121 genes were present in all nine tested strains, and 311 (6.1%) were K-12 specific and absent in the nine strains (Table 3). The remaining 580 genes (11.3%) were present in some strains and clearly absent or highly divergent in others. These variably absent or present genes we refer to as VAP genes. The majority of the VAP probes targeted prophage or phage-like genes. This was particularly striking for the Sakai-specific regions, where > 40% of all 1,084 Sakai-specific probes were VAP prophage and phage-like elements. In contrast, there are only 26 VAP elements (0.7%) in the 3,643 probes of the Sakai-K-12 backbone (Table 3). The percentage of Sakai-specific genes that are VAP genes is an order of magnitude greater (7.4%) than that for backbone genes (0.7%) but is 1/10 of that for phage-related Sakai genes (74.1%).

In order to analyze the presence and absence of genes in a phylogenetic framework, we combined the GACK analysis for the nine strains above with information inferred from the complete genomes of Sakai, EDL-933, and K-12. Over all 12 genomes, 3,663 (71.5%) of the genes are conserved, with the remaining 1,458 (38.5%) being VAP genes. There are 579 informative genes (variable targets present in at least two strains), and a PAUP (34) phylogenetic analysis of informative genes identified two maximum-parsimony trees (tree length = 839; homoplasy index = 0.310; retention index = 0.846), both with topologies consistent with the stepwise-evolution model. The phylogeny (Fig. 1B) based on all 1,458 VAP genes requires a total of 1,718 gains or losses of genes (839 steps plus 879 changes for noninformative genes), with most changes occurring between ancestral states A1 to A5. The parsimony analyses assume that gains or losses occur independently, which is not the case in nature for linked genes mobilized by phages or other mobile genetic elements.

SNPs. We assessed the level of sequence divergence by comparative sequencing of ~500 bp within each of 15 housekeeping genes (7,470 bp total), which uncovered seven single-nucleotide polymorphisms (SNPs) among the 11 genomes (K-12 excluded) (Table 4). Four of the seven SNPs are synonymous mutations, and three predict amino acid replacements. The comparison of the SNP distributions among strains (Table 4) and the stepwise model allow one to infer the step in which the point mutation occurred (Fig. 1B). For example, the mutation A776G in *uidA* (SNP-1) was inferred to have occurred between A3 and A5 previously by Monday et al. (24). In this way, mutations underlying all seven SNPs can be placed on the genome phylogeny as single events without any homoplasies (i.e., multiple origins, reversals, or losses).

To scale the sequence polymorphism to the variability in gene content, we calculated the genomic divergence in two ways. Based on the sequence data for 15 housekeeping genes,

TABLE 4. SNPs among *E. coli* O157:H7 and closely related strains

SNP	Locus	Position	Nucleotide in strain:											
			K-12	DEC 5d	TB182A	5905	493/89	CB-2755	G5101	ST-530	93-111	86-24	EDL-933	Sakai
1	*uidA*	776	A	A	A	A	A	A	G	G	G	G	G	G
2	*aspC*	267	G	G	G	G	A	A	G	G	G	G	G	G
3	*fadD*	1198	C	C	C	C	C	C	C	C	T	T	T	T
4	*rpoS*	543	C	C	C	C	C	C	C	C	A	A	A	A
5	*rpoS*	562	A	A	A	A	C	C	A	A	A	A	A	A
6	*cyaA*	528	C	C	C	T	C	C	C	C	C	C	C	C
7	*lysP*	169	C	T	T	T	T	T	T	T	T	T	C	T

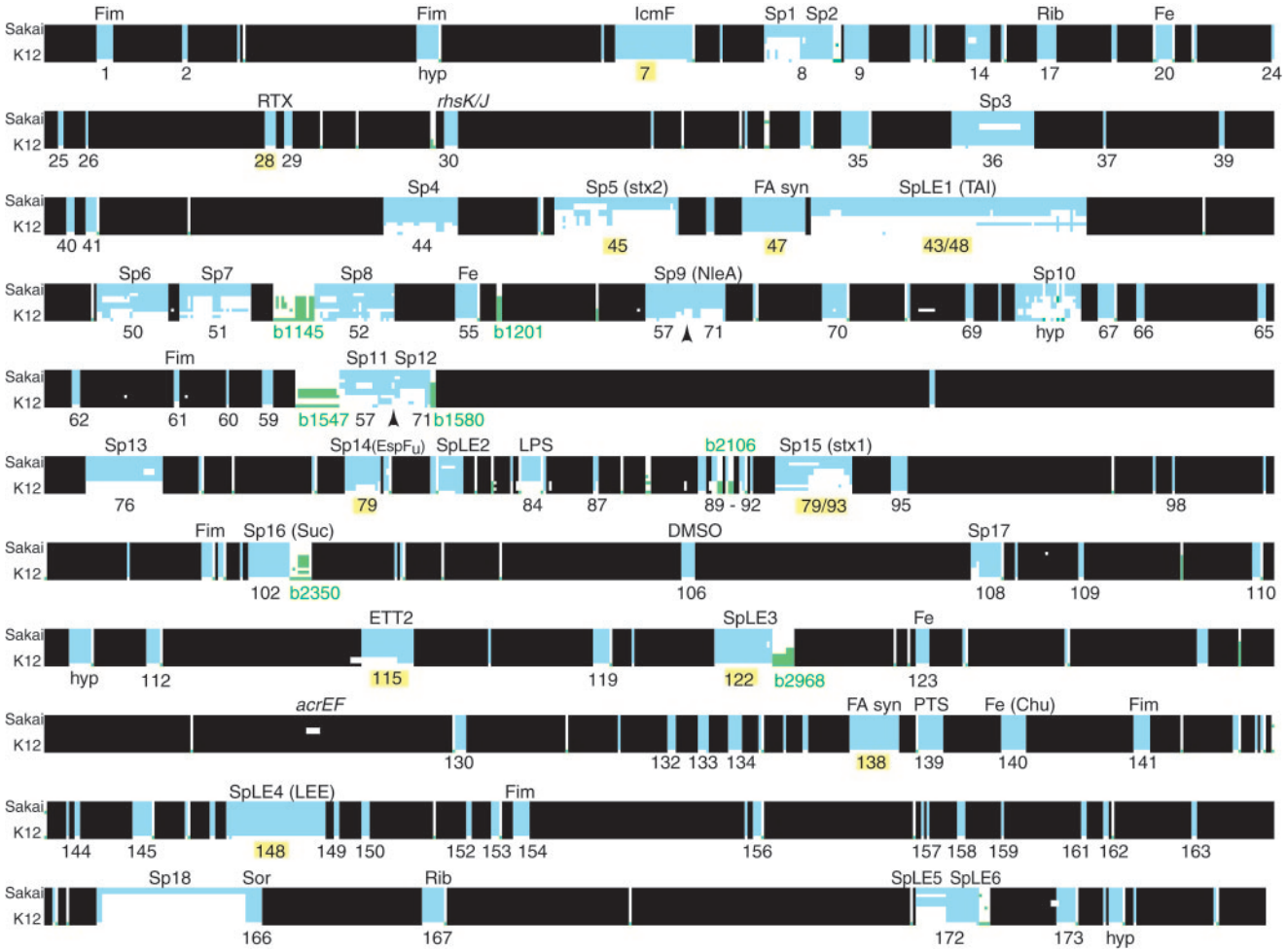

FIG. 3. Gene contents of strains from the stepwise evolution model for the emergence of *E. coli* O157:H7. Genes are ordered by their position in the Sakai genome from left to right and top to bottom. Genes are color coded as follows: black, backbone; blue, Sakai specific; green, K-12 specific. Rows represent 12 genomes in the phylogenetic order as shown in Fig. 1B: 1, Sakai; 2, 93–111; 3, EDL-933; 4, 86-24; 5, ST-530; 6, G5101; 7, CB2755; 8, 493/89; 9, 5905; 10, TB182A; 11, DEC 5d; 12, K-12. O islands are labeled by number below the rows, with those implicated in virulence highlighted in yellow. Known islands or putative functions are labeled above the rows (Fe, iron transport; Fim, fimbrial operon; FA syn, fatty acid synthesis; Rib, ribose transport; Sor, sorbose transport; Suc, sucrose transport). Arrowheads in Sp9 and Sp11/12 mark the sites of the large inversion in EDL-933 compared to Sakai.

the average nucleotide distance for a pair of genomes is 0.035% ± 0.143%. This is a measure of the degree of divergence between genomes resulting from the accumulation of point mutations in the time since the strains split from their most recent ancestor. In comparison, the pairwise genomic divergence in terms of the fraction of shared genes is 4.9% ± 0.18%. This measures the extent to which genes have been gained by horizontal transfer or duplication or lost by deletion in the same divergence time. The ratio of these two values roughly indicates that differences in gene content have accumulated at a rate ~140 times that for point mutations in housekeeping genes.

Islands. To investigate the genomic distribution of the presence or absence of polymorphisms, we plotted the binary data for 5,121 genes against map position in the Sakai genome (Fig. 3). Stretches of K-12-specific DNA were reduced for illustration (Fig. 3). Most (27 out of 33) of the larger (>5-kb) O islands identified in the comparison of the EDL-933 and K-12 genomes are conserved in gene content among the recent relatives of *E. coli* O157:H7. There are 27 variable regions where gain or loss of two or more adjacent genes has occurred (Fig. 3). Of the 18 Sakai prophages (Sp), only Sp16 (OI-102) was conserved in all strains, but only two of the six Sakai phage-like elements (SpLE), SpLE1 (TAI, OI-43/48) and SpLE 5 (OI-172) were variable.

For most prophages, subsets of genes, rather than the whole Sakai repertoire, are conserved among strains, suggesting frequent gains and losses of related phages or phage conversions in the recent past. Of the eight variable regions that are not phage related, five specify products that are cell surface related, including the O-antigenic region (OI-84), fimbriae, and adhesins. The remaining three regions include a secondary type III secretion system (OI-115, ETT2), an acridine resistance gene cluster (see "GUD$^+$ O157 strains" below), and a putative carbohydrate transport and metabolism gene cluster (ecs1890 to -95), which was divergent in O55 strain 5905.

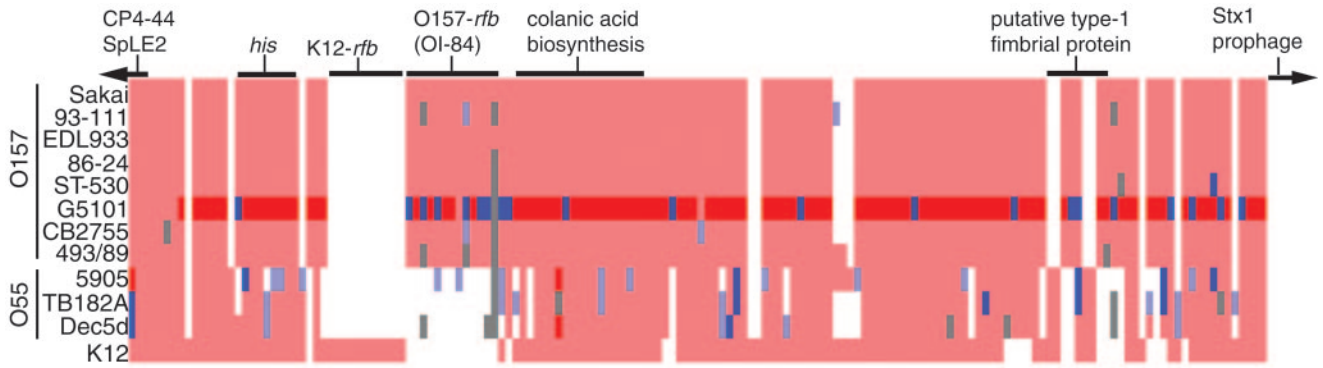

FIG. 4. Gene content in the region between SpLE2 and Stx1 prophage (ecs2806 to -2937). Genes are ordered by their position in the Sakai genome. Colors: white, absent; light red, present; dark red, duplicated; light blue, found present in one but absent in the other of two replicates (i.e., high probability of sequence divergence); dark blue, found duplicated in one of two replicates; gray, flagged spot in at least one of the replicates. Note that in addition to the genes determining the O antigens (*rfb*), several other genes in this region show a distinct pattern between O157 and O55 strains.

Variable regions of K-12-specific genes, inferred both by the direct and indirect methods (see Materials and Methods), identified nine regions (Fig. 3, with the b- number given for first gene in region), four of which are phage related and five of which have other putative functions. None of these regions were found in typical O157:H7 strains. Three regions were detected in all strains except the typical *E. coli* O157:H7; these are ORFs with unknown functions (b1160 to -72), genes for a putative sensor-type regulator and a putative adherence and penetration protein (b1201 to -2), and genes for starvation-sensing proteins (b1580 to -81). The other six regions found in some genomes contained secretion pathway and glycolate metabolism genes (b2968 to -86), fimbrial protein genes (b2106, b2108, b2111, and b2112), genes for cold shock-like proteins (b1550 to -68 on qin), prophage P2 proteins (b2082 to -84), and genes with unknown function (b1145 to -48 on e14 and b2356 to -61 on KpLE1).

LEE and Shiga toxins. The expression of both the LEE (SpLE4, OI-148) and the Shiga toxin genes contribute to the full virulence of EHEC strains. In agreement with previous results, the LEE has been found complete and intact in the *E. coli* O55:H7 strains and other close relatives of *E. coli* O157:H7. However, we found that *nleA*, which is located outside the LEE on Sp9, is absent in the O55:H7 genomes but present in the EHEC O157 strains. NleA is an effector secreted by the LEE-encoded type III secretion system and plays a key role in virulence of *Citrobacter rodentium* in a mouse infection model (10). $EspF_u$, encoded on Sp14, is a second effector secreted by EHEC that is not encoded by the LEE and has recently been shown to be essential for pedestal formation together with the LEE (3). We found the $EspF_u$ gene to be present in all of the close relatives of EHEC O157 examined here, suggesting that it was present in the progenitor of *E. coli* O55:H77 and EHEC O157:H7.

Microarray hybridizations correctly identified the Stx1 and Stx2 genes in toxin-producing strains. However, the presence of the toxin genes does not necessarily mean that the corresponding Stx-carrying Sakai phages (Sp5 for Stx2 and Sp15 for Stx1) have been conserved. In fact, all of the Stx1- or Stx2-producing strains before the proposed emergence of A6 (Fig. 1) are divergent in most phage genes found in Sp15 or Sp5 (Fig. 3). Both Sp5 (OI-45) and Sp15 (OI-79/93) were completed as recently as in the step leading to A6. The entire phage (all Sp5 genes) is missing in the Stx2-negative O55:H7 strains (DEC 5d and TB182A), whereas in the five Stx2-positive strains, only ~12 to 23% of Sp5 ORFs are found. Even among the typical O157:H7 strains EDL-933 and 86-24, there are stretches of >20 ORFs missing (Fig. 3, Sp5).

Strains lacking Stx1 contain several genes of the Sp15 phage. Sp15 in Sakai corresponds to OI-93, which is the Stx1 island, and parts of OI-79, which is located ~200 ORFs away from OI-93 in the EDL-933 genome. Interestingly, all strains except O55:H7 DEC 5d and O157:H7 86-24 contain the genes common to Sp15 and OI-79 (Fig. 3, left half of Sp15) but lack most of the ORFs common to Sp15 and OI-93 (Fig. 3, right half of Sp15). Even the Stx1-positive strains G5101 and ST-530 have only about 30% of the OI-93 genes.

O-antigen region. As expected, the complex locus specifying O157 lipopolysaccharide (OI-84) was absent in the three O55 strains and present in all O157 strains (Fig. 3 and 4). Previous sequence analysis of the *gnd* locus demonstrated an allelic difference between the O55:H7 and O157:H7 strains that led to the conclusion that *gnd* cotransferred during the antigenic shift (36). Wang et al. (39) identified a recombination site far downstream based on sequence differences in the *his* operon between strains O55 TB182A and Sakai. Interestingly, we found the same pattern of divergence between O55 and O157 strains for targets scattered over a stretch of about 100 ORFs between SpLE2 and Sp15 (ecs2819 to 2925) (Fig. 4). In this region, it appears that the K-12-specific fimbrial genes (b2106, b2108, b2111, and b2112) in the ancestral O55 strains were replaced by type 1 fimbrial genes (Fig. 4) in immediate O157 ancestors. In the O157 strain G5101, 103 of the127 targets between ecs2813 and ecs2937 were scored as duplications in replicate hybridization experiments (Fig. 4) suggesting the hypothesis that the whole region is duplicated in strain G5101. Together, these observations suggest that the entire 140-kbp segment, including three regions encoding surface properties (LPS, colanic acid biosynthesis, and type 1 fimbrial genes), was cotransferred horizontally and recombined in the O55-to-O157 antigenic shift.

Putative virulence factors. Comparison of the *E. coli* O157:H7 and K-12 genomes identified seven islands, in addition to the LEE, that are >15 kb in length and encode putative virulence factors (OI-7, OI-28, OI-47, OI-122, OI-138, OI-43/48, and OI-115) (31). The microarray hybridizations demonstrated that five of these islands (OI-7, OI-28, OI-47, OI-122, and OI-138) are conserved in the O157 lineage (Fig. 3). However, both GUD$^+$ O157 strains are highly divergent or have lost a putative EHEC adherence factor similar to Efa1 (27) within OI-122.

TAI, the tellurite resistance and adherence-conferring island (SpLE1), appears in O157 genomes after A3 and is duplicated in EDL-933 (OI-43 and OI-48). Interestingly, O55 strain 5905 contains about half of the island (ecs1359 to -1409), which could be a remnant or most likely was acquired independently. This part of the island contains the Iha adhesin but lacks the tellurite resistance and urease gene clusters. The GUD$^+$ strains are also highly divergent in two regions of the TAI island (ecs1306 to -1313 and ecs1384 to -1396), which encode an AIDA-1 adhesin-like protein (ecs1396) and a putative complement resistance protein (ecs1312).

The complete OI-115 island, a putative type III secretion system (ETT2), was found in all strains except the most ancestral O55:H7 strains (DEC 5d and TB182A). In these two strains at the base of the phylogeny, only about 25% of the island is present (ecs3731 to -3736), and the rest of the island (ecs3716 to -3730) and also a part of the backbone (ecs3709 to -3715) are missing. This pattern of loss suggests that the whole island was present in the common ancestor (A1 in Fig. 1) but has eroded as part of it, together with some backbone genes, were deleted in the lineage leading to DEC 5d and TB182A.

In addition to the factors on the LEE island and the Stx phages, 28 putative adhesin and toxin genes are found on the Sakai chromosome (11, 30, 31). Eleven of these are located on the seven large islands discussed above. The other 17 are conserved in all strains, with the exception of an intimin-like ORF on the OI-173 (ecs5290) and two other putative adhesin genes (ecs0350 on OI-14 and ecs2776) that are divergent in the GUD$^+$ O157 strains. Finally, 12 of the 14 loci for fimbrial biosynthesis (11) were found in all strains. One fimbrial locus showed the same pattern of distribution as the O-antigenic region (see above), and one locus was present in all except TB182A (ecs2112 to -2113). It has been shown recently that the latter fimbrial locus (ecs2113) is important in colonization of calves (5); however, its role in human disease is unknown. All strains also contain the six iron uptake systems found in Sakai and are missing the *fec* transport system found in K-12.

GUD$^+$ O157 strains. The genomes of the GUD$^+$ O157 strains showed several stretches of divergence scattered all over the chromosome. As noted above, these O157 strains are highly divergent in islands encoding five putative adherence factors and, in addition, have lost or are highly divergent in at least 10 more genomic regions, many of which contain genes of unknown function. One lost region specifies a putative complement resistance protein, TraT, and two others contain genes for resistance against acridine (ecs4134 to -4139) and methylviologen (ecs1611 to -1619). Clearly, none of the deleted regions are absolutely required for virulence of O157 strains.

DISCUSSION

Comparative genomic hybridization with 50-mer oligoarrays reliably differentiated *E. coli* Sakai and K-12, whose genome sequences are known. Using GACK analysis optimized to infer the presence and absence of genes, we estimate that 50% of the probes with divergent targets of ~94% sequence identity are scored as absent. A comparable GACK analysis of a PCR-based *Helicobacter pylori* array assigned 50% of genes with ~89% identity as divergent (18). The majority of PCR products on the *H. pylori* array had 93 to 97% sequence identity to the test strain (18), whereas the majority of probes on the O157 oligoarray are 100% identical to the test genome targets. Because this majority defines gene presence, in a test strain with most sequences slightly divergent from the reference strain (as in the *H. pylori* case), the sequence identity value below which genes are called absent or highly divergent will also be lower. Thus, the detection limit of oligoarrays is nearly equivalent to that of PCR-based whole-ORF arrays when investigating present or absent polymorphisms in closely related genomes.

The hybridization data are consistent with previous knowledge about the mobile virulence elements in the pathogens investigated here, which further validates the accuracy of these multigenome oligoarrays. The presence of the Stx1 and Stx2 genes and the LEE island was correctly assigned in all tests. Our analysis showed that the whole SpLE5 island (half of OI-172) is missing in O157 strain 86-24, a result consistent with a previous study (20). The results also indicate that tellurite resistance encoded by the TAI island was acquired recently in the step from A3 to A5, as proposed by Tarr et al. (35). The observation that *stx1* is missing but elements of the Stx1 phage are present in 86-24 (Sp15 in Fig. 3) was also reported previously (33).

The Stx1 and Stx2 phages have a complex history even over the short time scale separating the immediate ancestors of *E. coli* O157:H7. Shaikh and Tarr (33) mapped the phage integration sites, which, with our results, indicate that the Stx1 and Stx2 phages integrated at the same site as in the Sakai strain (*wrbA* for Stx2 and *yehV* for Stx1) are similar in gene content, whereas phages integrated into other sites in related strains are divergent in gene content. Data from Ohnishi et al. (29), who determined position and diversity of Stx phages in eight O157 strains from Japan by PCR scanning, showed the same pattern. Together these findings support a scenario in which the phage-borne toxin genes were acquired early and conserved despite evidence of dynamic turnover in phage genes, resulting from phage replacement, localized recombination, and island erosion. The comparative genomic analysis indicates that after occupation of *yehV* by the Stx1 phage (or truncated Stx1 phage) and the occupation of *wrbA* by the Stx2 phage, these prophages diversified and gained the additional genes (or the prophages were replaced or recombined with other phages) to achieve the gene complement of O157 Sakai.

We identified the Sakai and K-12 genes that were gained or lost during the emergence of *E. coli* O157:H7 from its O55:H7-like ancestor. Overall, the phylogeny inferred from differences in gene content confirms the stepwise evolution model (7). The phylogeny shows that the representative intermediates have diverged from the proposed ancestral nodes (A1 to A6). It also reveals that the variation between pairs of strains de-

rived from each ancestral node is small. For example, the two typical *E. coli* O55:H7 strains (DEC 5d and TB182A) descended from A1 are very similar but are clearly different from the atypical Stx2-positive O55:H7 strain (5905) derived from A2. The multigenome array used here also permits a second, independent enumeration of present or absent polymorphisms by using only genes present in K-12, excluding Sakai-specific genes. This second tree topology is identical to that in Fig. 1B with the exception that K-12 clusters with the O157 German clone, which contains several K-12-specific genes (Fig. 3) (b1145, b1547, and b2350) that are absent in most other strains. These results demonstrate that a K-12 array is useful not only for phylogenetic analysis of pathogenic *E. coli* strains that are distantly related among each other, as shown by Fukiya et al. (8), but even for phylogenetic resolution of a group of closely related strains.

About 85% of the genes that are variably present (so-called VAP genes) are phage related, underscoring the dominant role of phages in diversification of the chromosomal architecture of *E. coli* O157 strains (11, 31). Several Sakai prophages vary in gene content even among very closely related host strains (Fig. 3), indicating that phage genomes themselves rapidly diversify. Thus, these bacteria can act as "phage factories," producing a variety of chimeric phages (28). The extent to which these differences in gene content among phages contribute to variation in virulence among toxin-producing strains has yet to be elucidated.

To investigate further the genomic impact of phage variability, we classified the VAP genes into clusters of orthologous proteins (COGs). Most VAP genes could not be classified. Of those that could be classified, >20% of the 218 genes involved in DNA replication, recombination, and repair are the most variable (Fig. 5). This is because most VAP genes are phage related, so that COGs with a high fraction of phage-related genes also showed a high percentage of VAP genes (correlation coefficient = 0.97, $P < 0.001$, $t = 17.43$). In fact, there is a significant excess of phage-related VAP genes in DNA replication and recombination (e.g., integrases) and transcription (Fig. 4). In contrast, there was a significant deficiency of phage-related VAP genes associated with proteins involved in secretion, cell motility, cell membrane biogenesis, and carbohydrate metabolism (Fig. 4). Overall, the most conserved COGs include genes involved in either nucleotide or coenzyme transport. Interestingly, the more stable groups among the phage-related genes seem to be the more variable groups among the non-phage-related genes. This inverse relationship suggests the hypothesis that most phage genes are gained and lost quickly, except for the ones conferring advantages to the bacterial host, which are retained by natural selection. The same selective pressure drives diversity in non-phage-related genes of these functional groups. It is also possible that the phage origins might be obscured after the erosion of the extraneous phage genes so that the remaining genes are considered to be native and not phage related in origin. Phage and cell envelope genes were identified as the ones most often horizontally transferred based on sequence analysis of 116 prokaryotic genome sequences (25).

Interestingly, most of the ancillary virulence factors identified after completion of the genomic sequences of two O157:H7 strains were already present in the O55:H7-like ancestor (A1 in Fig. 1). It remains to be elucidated which of the factors gained or lost since then made today's O157:H7 strains such potent pathogens. For example, $EspF_u$ (on Sp14) was found in all strains, whereas NleA (on Sp9) was absent in the O55:H7 strains but present in the O157 strains. Besides factors directly involved in virulence, general fitness factors contribute to the ability of *E. coli* O157:H7 to survive in environmental niches; this survival enhances the chance to infect a host and thus also may play an important and indirect role in determining the pathogenicity and epidemiology of a pathogenic clone. Point mutations in functional or regulatory genes, which can also have such effects, are normally not detected by comparative genomic hybridization analysis, as are losses and gains of genes not represented on the chip. Two examples of mutations detectable in phenotypic assays but not with the array used in this study are the loss of abilities to ferment sorbitol (SOR^-) and to express β-glucuronidase activity (GUD^-). GUD^- strains carry the *uidA* gene (which encodes GUD), but a frameshift in the gene prevents expression of a functional protein (24). The SOR^- phenotype also is probably caused by frameshifts that are present in the *srlA* and *srlE* genes of the Sakai and the EDL-933 strains (11, 31). In K-12, the intact *srlA* and *srlE* encode components of a glucitol/sorbitol-specific phosphotransferase system (2).

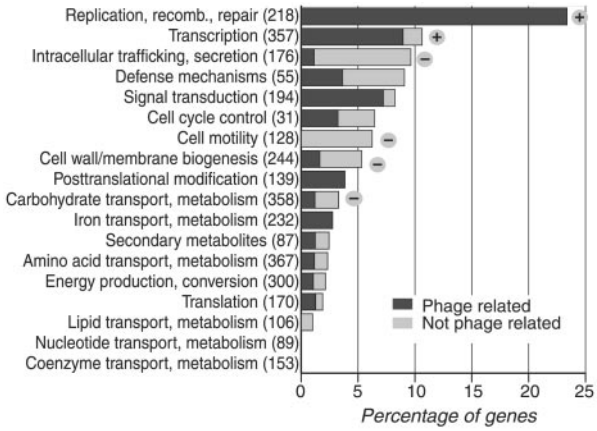

FIG. 5. Percentages of VAP genes in COGs. For each functional group, the fraction of genes that are VAP genes is shown, divided into phage-related and non-phage-related categories. Functional groups were defined by the COG database (37). Significant excesses (+) and deficiencies (−) of phage-related VAP genes were determined by residual analysis of the contingency table (6).

Surprisingly, the GUD^+ O157 strains, although phylogenetically closest to the typical O157:H7 strains, are highly divergent in several loci encoding adherence factors and defense mechanisms; such changes in adherence properties can indicate shifts in the environmental niche of a bacterium (13). Thus, it appears that genomic dynamics, primarily fostered in the O157 lineage by phage mobility, island acquisitions, and subsequent erosions, is a trial-and-error process that can promote genetic change underlying the ecological diversification of bacterial pathogens.

ACKNOWLEDGMENTS

We thank Lindsey Ouellette for technical assistance and James Rudrick of the Michigan Department of Community Health for sup-

plying *E. coli* strain ST-530. We also appreciate Phillip Tarr and Peter Feng for reviewing a previous version of the manuscript.

This project has been funded in part by the MSU foundation and in part with funds from NIH grant AI47499. The STEC Center is supported with funds from the NIAID, NIH, DHHS, under NIH research contract N01-AI-30058.

REFERENCES

1. **Bilban, M., L. Buehler, S. Head, G. Desoye, and V. Quaranta.** 2002. Defining signal thresholds in DNA microarrays: exemplary application for invasive cancer. BMC Genomics **3:**19.
2. **Blattner, F., G. Plunkett III, C. Bloch, N. Perna, V. Burland, M. Riley, J. Collado-Vides, J. Glasner, C. Rode, G. Mayhew, J. Gregor, N. Davis, H. Kirkpatrick, M. Goeden, D. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.

2a. **Bokete, T. N., T. S. Whittam, R. A. Wilson, C. R. Clausen, C. M. O'Callahan, S. L. Moseley, T. R. Fritsche, and P. L. Tarr.** 1997. Genetic and phenotypic analysis of *Escherichia coli* with enteropathogenic characteristics isolated from Seattle children. J. Infect. Dis. **175:**1382–1389.

3. **Campellone, K. G., D. Robbins, and J. M. Leong.** 2004. $EspF_U$ is a translocated EHEC effector that interacts with Tir and N-WASP and promotes Nck-independent actin assembly. Dev. Cell **7:**217–228.
4. **Donnenberg, M. S., and T. S. Whittam.** 2001. Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. J. Clin. Investig. **107:**539–548.
5. **Dziva, F., P. M. van Diemen, M. P. Stevens, A. J. Smith, and T. S. Wallis.** 2004. Identification of *Escherichia coli* O157:H7 genes influencing colonization of the bovine gastrointestinal tract using signature-tagged mutagenesis. Microbiology **150:**3631–3645.
6. **Everitt, B. S.** 1977. The analysis of contingency tables. Chapman and Hall, London, United Kingdom.
7. **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.
8. **Fukiya, S., H. Mizoguchi, T. Tobe, and H. Mori.** 2004. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. J. Bacteriol. **186:**3911–3921.
9. **Griffin, P. M.** 1995. *Escherichia coli* O157:H7 and other enterohemorrhagic *Escherichia coli*, p. 739–761. *In* M. J. Blaser, P. D. Smith, J. I. Ravdin, H. B. Greenberg, and R. L. Guerrant (ed.), Infections of the gasterointestinal tract. Raven Press, New York, N.Y.
10. **Gruenheid, S., I. Sekirov, N. A. Thomas, W. Deng, P. O'Donnell, D. Goode, Y. Li, E. A. Frey, N. F. Brown, P. Metalnikov, T. Pawson, K. Ashman, and B. B. Finlay.** 2004. Identification and characterization of NleA, a non-LEE-encoded type III translocated virulence factor of enterohaemorrhagic *Escherichia coli* O157:H7. Mol. Microbiol. **51:**1233–1249.
11. **Hayashi, T., K. Makino, M. Ohnishi, K. Kurokawa, K. Ishii, K. Yokoyama, C. G. Han, E. Ohtsubo, K. Nakayama, T. Murata, M. Tanaka, T. Tobe, T. Iida, H. Takami, T. Honda, C. Sasakawa, 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 K-12. DNA Res. **8:**11–22.

11a. **Hayes, P. S., K. Blorn, P. Feng, J. Lewis, N. A. Strockbine, and B. Swaminathan.** 1995. Isolation and characterization of a β-D-glucuronidase-producing strain of *Escherichia coli* serotype O157:H7 in the United States. J. Clin. Microbiol. **33:**3347–3348.

12. **Hyma, K. E., D. W. Lacher, A. M. Nelson, A. C. Bumbaugh, J. M. Janda, N. A. Strockbine, V. B. Young, and T. S. Whittam.** 2005. Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. J. Bacteriol. **187:**619–628.
13. **Johnson, J. R.** 2002. Evolution of pathogenic *Escherichia coli*, p. 55–77. *In* M. S. Donnenberg (ed.), *Escherichia coli*: virulence mechanisms of a versatile pathogen. Academic Press, San Diego, Calif.
14. **Kaper, J. B., J. P. Nataro, and H. L. T. Mobley.** 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. **2:**123–140.
15. **Karch, H., and M. Bielaszewska.** 2001. Sorbitol-fermenting Shiga toxin-producing *Escherichia coli* O157:H− strains: epidemiology, phenotypic and molecular characteristics, and microbiological diagnosis. J. Clin. Microbiol. **39:**2043–2049.
16. **Karch, H., M. Bielaszewska, M. Bitzan, and H. Schmidt.** 1999. Epidemiology and diagnosis of Shiga toxin-producing *Escherichia coli* infections. Diagn. Microbiol. Infect. Dis. **34:**229–243.
17. **Karch, H., H. Bohm, 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.
18. **Kim, C., E. Joyce, K. Chan, and S. Falkow.** 2002. Improved analytical methods for microarray-based genome-composition analysis. Genome Biol. **3:**Research0065.1–Research0065.17.
19. **Kim, J., J. Nietfeldt, J. Ju, J. Wise, N. Fegan, P. Desmarchelier, and A. K. Benson.** 2001. Ancestral divergence, genome diversification, and phylogeographic variation in subpopulations of sorbitol-negative, beta-glucuronidase-negative enterohemorrhagic *Escherichia coli* O157. J. Bacteriol. **183:**6885–6897.
20. **Kudva, I. T., R. W. Griffin, M. Murray, M. John, N. T. Perna, T. J. Barrett, and S. B. Calderwood.** 2004. Insertions, deletions, and single-nucleotide polymorphisms at rare restriction enzyme sites enhance discriminatory power of polymorphic amplified typing sequences, a novel strain typing system for *Escherichia coli* O157:H7. J. Clin. Microbiol. **42:**2388–2397.
21. **Kumar, S., K. Tamura, I. Jakobsen, and M. Nei.** 2000. MEGA 2: Molecular Evolutionary Genetics Analysis program, version 2.0. Pennsylvania State University, University Park.
22. **Mead, P. S., and P. M. Griffin.** 1998. *Escherichia coli* O157:H7. Lancet **352:**1207–1212.
23. **Monday, S. R., S. A. Minnich, and P. C. Feng.** 2004. A 12-base-pair deletion in the flagellar master control gene *flhC* causes nonmotility of the pathogenic German sorbitol-fermenting *Escherichia coli* O157:H− strains. J. Bacteriol. **186:**2319–2327.
24. **Monday, S. R., T. S. Whittam, and P. C. Feng.** 2001. Genetic and evolutionary analysis of mutations in the *gusA* gene that cause the absence of beta-glucuronidase activity in *Escherichia coli* O157:H7. J. Infect. Dis. **184:**918–921.
25. **Nakamura, Y., T. Itoh, H. Matsuda, and T. Gojobori.** 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat. Genet. **36:**760–766.
26. **Nataro, J. P., and J. B. Kaper.** 1998. Diarrheagenic *Escherichia coli*. Clin. Microbiol. Rev. **11:**142–201.
27. **Nicholls, L., T. H. Grant, and R. M. Robins-Browne.** 2000. Identification of a novel genetic locus that is required for in vitro adhesion of a clinical isolate of enterohaemorrhagic *Escherichia coli* to epithelial cells. Mol. Microbiol. **35:**275–288.
28. **Ohnishi, M., K. Kurokawa, and T. Hayashi.** 2001. Diversification of *Escherichia coli* genomes: are bacteriophages the major contributors? Trends Microbiol. **9:**481–485.
29. **Ohnishi, M., J. Terajima, K. Kurokawa, K. Nakayama, T. Murata, K. Tamura, Y. Ogura, H. Watanabe, and T. Hayashi.** 2002. Genomic diversity of enterohemorrhagic *Escherichia coli* O157 revealed by whole genome PCR scanning. Proc. Natl. Acad. Sci. USA **99:**17043–17048.
30. **Perna, N. T., J. D. Glasner, V. Burland, and G. Plunkett III.** 2002. The genomes of *Escherichia coli* K12 and pathogenic *E. coli*, p. 3–53. *In* M. S. Donnenberg (ed.), *Escherichia coli*: virulence mechanisms of a versatile pathogen. Academic Press, San Diego, Calif.
31. **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. Potamousis, J. Apodaca, T. S. Anantharaman, J. Lin, G. Yen, D. C. Schwartz, R. A. Welch, and F. R. Blattner.** 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature **409:**529–533.
32. **Reid, S. D., C. J. Herbelin, A. C. Bumbaugh, R. K. Selander, and T. S. Whittam.** Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature **406:**64–67.
33. **Shaikh, N., and P. I. Tarr.** 2003. *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications. J. Bacteriol. **185:**3596–3605.
34. **Swofford, D. L.** 2000. PAUP and other methods. Phylogenetic analysis using parsimony, 4th ed. Sinauer Associates, Sunderland, Mass.
35. **Tarr, P. I., S. S. Bilge, J. C. Vary, Jr., S. Jelacic, R. L. 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.

35a. **Tarr, P. I., M. A. Neill, C. R. Clausen, J. W. Newland, R. J. Neill, and S. L. Moseley.** 1989. Genotypic variation in pathogenic *Escherichia coli* O157:H7 isolated from patients in Washington, 1984–1987. J. Infect. Dis. **159:**344–347.

36. **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:**6183–6191.
37. **Tatusov, R., N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale.** 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4:**41.
38. **Tatusova, T., I. Karsch-Mizrachi, and J. Ostell.** 1999. Complete genomes in WWW Entrez: data representation and analysis. Bioinformatics **15:**536–543.
39. **Wang, L., S. Huskic, A. Cisterne, D. Rothemund, and P. R. Reeves.** 2002. The O-antigen gene cluster of *Escherichia coli* O55:H7 and identification of a new UDP-GlcNAc C4 epimerase gene. J. Bacteriol. **184:**2620–2625.
40. **Whittam, T. S., M. L. Wolfe, I. K. Wachsmuth, F. Ørskov, I. Ørskov, and R. A. Wilson.** 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect. Immun. **61:**1619–1629.
41. **Wu, H., K. M. Kerr, X. Q. Cui, and G. A. Churchill.** 2003. MAANOVA: a software package for the analysis of spotted cDNA microarray experiments. *In* G. Parmigiani, E. S. Garret, R. A. Irizarry, and S. L. Zeger (ed.), Analysis of gene expression data: an overview of methods and software. Springer, New York, N.Y.