水口雅, 種市尋宙.	EHEC 感染症による脳症の治療.	五十嵐隆 (総括	溶血性尿毒症症候群の診断・治療ガイドライン	東京医学社	東京	2014	50-56

別紙 5-1

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR.	A complete view of the genetic diversity of the Escherichia coli O-antigen bi osynthesis gene cluster.	DNA Research	22	101-107	2015
von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A, Thomson NR, Rasko DA, Joffre E, Corander J, Pickard D, Wiklund G, Svennerholm AM, Sjöling Å, Dougan G.	Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution.	Nature Genetics	46	1321-1326	2014
Mekata H, Iguchi A, Kawano K, Kirino Y, Kobayashi I, Misawa N.	Identification of O serotypes, genotypes, and virulotypes of Shiga toxin-producing Escherichia coli isolates, including non-O157 from beef cattle in Japan.	Journal of Food Protection	77	1269-1274	2014
Iyoda S, Manning SD, Seto K, Kimata K, Isobe J, Etoh Y, Ichihara S, Migita Y, Ogata K, Honda M, Kubota T, Kawano K, Matsumoto K, Kudaka J, Asai N, Yabata J, Tominaga K, Terajima J, Morita-Ishihara T, Izumiya H, Ogura Y, Saitoh T, Iguchi A, Kobayashi H, Hara-Kudo Y, Ohnishi M, EHEC working group in Japan	Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular *stx* subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers.	Open Forum Infectious Disease	1		2014

N. Sudo, A. Soma, A. Muto, S. Iyoda, M. Suh, N. Kurihara, H. Abe, T. Tobe, Y. Ogura, T. Hayashi, K. Kurokawa, M. Ohnishi, Y. Sekine	A novel small regulatory RNA accelerates cell motility in enterohemorrhagic *Escherichia coli.*	J. Gen. Appl. Microbiol.	60	44-50	2014
M. Kusumoto, D. Fukamizu, Y. Ogura, E. Yoshida, F. Yamamoto, T. Iwata, T. Ooka, M. Akiba, T. Hayashi	The lineage-specific distribution of IS-excision enhancer in enterotoxigenic *Escherichia coli* isolated from swine	Appl. Environ. Microbiol.	80(4)	1394-1402	2014
A. Hinenoya, K. Shima, M. Asakura, K. Nishimura, T. Tsukamoto, T. Ooka, T. Hayashi, T. Ramamurthy, S. Faruque and S. Yamasaki	Molecular characterization of cytolethal distending toxin gene-positive *Escherichia coli* from healthy cattle and swine in Nara, Japan.	BMC Microbiology	14	97	2014
R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, T. Itoh	Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.	Genome Res.	24(8)	1384-1395	2014
A. Iguchi, S. Iyoda, T. Kikuchi, Y. Ogura, K. Katsura, M. Ohnishi, T. Hayashi, N. R. Thomson	A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster.	DNA Res.	22(1)	101-107	2015
Watahiki, M., Isobe, J., Kimata, K., Shima, T., Kanatani, J., Shimizu, M., Nagata, A., Kawakami, K., Yamada, M., Izumiya, H.,	Characterization of enterohemorrhagic *Escherichia coli* 0111 and 0157 strains isolated from outbreak patients in Japan	Journal of Clinical Microbiology	Vol. 52 No. 8	2757-2763	2014

Iyoda, S., Morita-Ishihara, T., Mitobe, J., Terajima, J., Ohnishi, M. and Sata, T.					
Yahiro, K., H. Tsutsuki, K. Ogura, S. Nagasawa, J. Moss, and M. Noda.	DAP1, a Negative Regulator of Autophagy, Controls SubAB-Mediated Apoptosis and Autophagy.	Infect Immun	82	4899-908	2014
Nagasawa, S., K. Ogura, H. Tsutsuki, H. Saitoh, J. Moss, H. Iwase, M. Noda, and K. Yahiro.	Uptake of Shiga-toxigenic Escherichia coli SubAB by HeLa cells requires an actin- and lipid raft-dependent pathway.	Cell Microbiol.	16	1582-601	2014
Igarashi T, Ito S, Sako M, Saitoh A, Hataya H, Mizuguchi M, Morishima T, Ohnishi K, Kawamura N, Kitayama H, Ashida A, Kaname S, Taneichi H, Tang J, Ohnishi M, Study group for establishing guidelines for the diagnosis and therapy of hemolytic uremic syndrome	Guidelines for the management and investigation of hemolytic uremic syndrome	Clin Exp Nephrol	18	525-557	2014
Hattori M, Sako M, Kaneko T, Ashida A, Matsunaga A, Igarashi T, Itami N, Ohta T, Gotoh Y, Satomura K, Honda M, Igarashi T	End-stage renal disease in Japanese children: a nationwide survey during 2005-2011	Clin Exp Nephrol	In press		2014

Sawai T, Nangaku M, Ashida A, Fujimaru R, Hataya H, Hidaka Y, Kaname S, Okada H, Sato W, Yasuda T, Yoshida Y, Fujimura Y, Hattori M, Kagami S	Diagnostic criteria for atypical hemolytic uremic syndrome proposed by the Joint Committee of the Japanese Society of Nephrology and the Japan Pediatric Society	Clin Exp Nephrol	18	4-9	2014
Sawai T, Nangaku M, Ashida A, Fujimaru R, Hataya H, Hidaka Y, Kaname S, Okada H, Sato W, Yasuda T, Yoshida Y, Fujimura Y, Hattori M, Kagami	Diagnostic criteria for atypical hemolytic uremic syndrome proposed by the Joint Committee of the Japanese Society of Nephrology and the Japan Pediatric Society	Pediatr Int	56	1-5	2014
芦田 明、玉井浩	溶血性尿毒症症候群（典型的/非典型的）	小児内科	46	209-213	2014
芦田 明、玉井浩	補体調節因子異常による非典型溶血性尿毒症症候群（aHUS）	小児科診療	77	771-777	2014
芦田 明、玉井浩	エクリズマブ：aHUS	腎と透析	76	77-80	2014
芦田 明、玉井浩:	典型的 HUS	日本血栓止血学会雑誌	25	706-712	2014

芦田　明、玉井浩	非典型溶血性尿毒症症候群	日本アフェレシス学会雑誌	34	40-47	2014
芦田　明、玉井浩	aHUS とアフェレシス	日本アフェレシス学会雑誌	34	40-47	2014
北山 浩嗣・和田尚弘	新生児・小児の敗血症に対するエンドトキシン吸着療法（PMX-DHP）	日本アフェレシス学会雑誌	34		2015 (in press)
種市 尋宙, 六車崇, 太田 邦雄, 小西 道雄, 奥村彰久, 高梨 潤一, 水口 雅, 宮脇 利男	腸管出血性大腸菌 O111 集団感染における危機対応	日本小児科学会雑誌	118	103-108	2014
種市尋宙	腸管出血性大腸菌 O111 集団感染における小児重症例の特徴とその対策	小児科	55	1017-1025	2014
種市尋宙	腸管出血性大腸菌 O111 の治療：集団感染から学ぶもの	感染症内科	2	195-202	2014

種市尋宙	腸管出血性大腸菌感染症による急性脳症の病態と治療戦略	Neuroinfection			2015 (in press)
Shimizu M, Kuroda M, Inoue N, Konishi M, Igarashi N, Taneichi H, Kanegane H, Ito M, Saito S, Yachie A	Extensive serum biomarker analysis in patients with enterohemorrhagic *Escherichia coli* O111-induced hemolytic-uremic syndrome.	Cytokine	66	1-6	2014
Takanashi J, Taneichi H, Misaki T, Yahata Y, Okumura A, Ishida Y, Miyawaki T, Okabe N, Sata T, Mizuguchi M	Clinical and radiologic features of encephalopathy during 2011 *E coli* O111 outbreak in Japan	Neurology	82	564-572	2014
Igarashi T, Ito S, Sako M, et al.	Study group for establishing guide lines for the diagnosis and therapy of hemolytic uremic syndrome: Guidelines for the management and investigation of hemolytic uremic syndrome.	Clin Exp Nephro	18	525-557	2014
（総括責任者：五十嵐 隆）伊藤秀一	溶血性尿毒症症候群の診断・治療ガイドライン	東京医学社			2014
Hattori M, Sako M, Kaneko T, Ito S. et al.	End-stage renal disease in Japanese children: a nationwide survey during 2005-2011.	Clin Exp Nephrol	(in press)		2014

伊藤　秀一	HUS up to date　欧州における大規模集団感染を中心に	日本小児腎不全学会雑誌	33	16-19	2013
伊藤　秀一	志賀毒産生性大腸菌によるHUSの治療	日本腎臓病学会雑誌	56（7）	1075-81	2014
Masaki Fuyama,Masaki Takahaahi,Shuichi Ito et al	Efficacy of cyclophosphamide and mizoribine combination therapy against steroid dependent nephrotic syndrome;	ACPN New Delhi(India)			2014
MasakiTakahashi, KoichiKamei,Shuichi Ito,et al	Analysis of risk factors for a cyclosporine nephrotoxicity in children with idiopathic nephrotic syndrome	ACPN New Delhi(India)			2014
KoichiKamei,IsaoMiyairi,Shuichi Ito,etal	Prospecive trial of attenuated live vaccines in children receiving immunosuppressants	ACPN New Delhi(India)			2014
伊藤　秀一	ステロイド抵抗性ネフローゼ症候群に対する他施設臨床試験	日本小児科学会総会・学術集会　名古屋（愛知）	118(2)	160-42	2014

伊藤　秀一	HUSRevisited	日本小児科学会総会・学術集会　名古屋（愛知）	118(2)	139-21	2014
Takanashi J, Taneichi H, Misaki T, Yahata Y, Okumura A, Ishida Y, Tanaka T, Miyawaki T, Okabe N, Mizuguchi M.	Clinical and radiological features of encephalopathy during 2011 *E. coli* O111 outbreak in Japan.	Neurology	82(7)	564-572	2014
種市尋宙，六車崇，太田邦雄，小西道雄，住田亮，奥村彰久，高梨潤一，水口雅，宮脇利男.	腸管出血性大腸菌 O111 集団感染における危機対応.	日本小児科学会雑誌	118(7)	1103-1108	2014
Igarashi T, Ito S, Sako M, Saitoh A, Hataya H, Mizuguchi M, Morishima T, Ohnishi K, Kawamura N, Kitayama H, Ashida A, Kaname S, Taneichi H, Tang J, Ohnishi M; Study group for establishing guidelines for the diagnosis and therapy of hemolytic uremic syndrome.	Guidelines for the management and investigation of hemolytic uremic syndrome.	Clinical and Experimental Nephrology	18(4)	525-557	2014

DNA Research, 2015, 22(1), 101–107
doi: 10.1093/dnares/dsu043
Advance Access Publication Date: 26 November 2014
Full Paper

Full Paper

A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster

Atsushi Iguchi[1,*], Sunao Iyoda[2], Taisei Kikuchi[3], Yoshitoshi Ogura[4,5], Keisuke Katsura[5], Makoto Ohnishi[2], Tetsuya Hayashi[4,5], and Nicholas R. Thomson[6,7]

[1]Department of Animal and Grassland Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki 889-2192, Japan, [2]Department of Bacteriology I, National Institute of Infectious Diseases, Tokyo 162-8640, Japan, [3]Division of Parasitology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [4]Division of Microbiology, Department of Infectious Diseases, Faculty of Medicine, University of Miyazaki, Miyazaki 889-1692, Japan, [5]Division of Bioenvironmental Science, Frontier Science Research Center, University of Miyazaki, Miyazaki 889-1692, Japan, [6]Pathogen Genomics, The Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK, and [7]Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK

*To whom correspondence should be addressed. Tel/Fax. +81 985-58-7507. E-mail: iguchi@med.miyazaki-u.ac.jp

Edited by Dr Katsumi Isono

Received 16 September 2014; Accepted 3 November 2014

Abstract

The O antigen constitutes the outermost part of the lipopolysaccharide layer in Gram-negative bacteria. The chemical composition and structure of the O antigen show high levels of variation even within a single species revealing itself as serological diversity. Here, we present a complete sequence set for the O-antigen biosynthesis gene clusters (O-AGCs) from all 184 recognized *Escherichia coli* O serogroups. By comparing these sequences, we identified 161 well-defined O-AGCs. Based on the *wzx*/*wzy* or *wzm*/*wzt* gene sequences, in addition to 145 singletons, 37 serogroups were placed into 16 groups. Furthermore, phylogenetic analysis of all the *E. coli* O-serogroup reference strains revealed that the nearly one-quarter of the 184 serogroups were found in the ST10 lineage, which may have a unique genetic background allowing a more successful exchange of O-AGCs. Our data provide a complete view of the genetic diversity of O-AGCs in *E. coli* showing a stronger association between host phylogenetic lineage and O-serogroup diversification than previously recognized. These data will be a valuable basis for developing a systematic molecular O-typing scheme that will allow traditional typing approaches to be linked to genomic exploration of *E. coli* diversity.

Key words: *E. coli*, O-antigen biosynthesis gene cluster, horizontal gene transfer, O serogroup, genomic diversity

1. Introduction

Cell-surface polysaccharides play an essential role in the ability of bacteria to survive and persist in the environment and in host organisms.[1] The O-antigen polysaccharide constitutes the outermost part of the lipopolysaccharide (LPS) present in the outer membrane of Gram-negative bacteria. The chemical composition and structure of the O-antigen exhibit high levels of variation even within a single species.[2–5] This observation is corroborated by the huge serological

© The Author 2014. Published by Oxford University Press on behalf of Kazusa DNA Research Institute. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

variation of somatic O antigens. Currently, the O serogrouping, sometimes combined with H (flagellar) antigens and K (capsular polysaccharide) antigens, is a standard method for subtyping of *Escherichia coli* strains in taxonomical and epidemiological studies. In particular, identification of strains of the same O serogroup is a prerequisite to start any actions for outbreak investigations and surveillance.

Thus far, the World Health Organization Collaborating Centre for Reference and Research on *Escherichia* and *Klebsiella* based at the Statens Serum Institut (SSI) in Denmark (http://www.ssi.dk/English.aspx) has recognized 184 *E. coli* O serogroups. It is generally believed that the O serogrouping of *E. coli* strains provides valuable information for identifying pathogenic clonal groups, especially for public health surveillance. For example, O157 is a leading O serogroup associated with enterohemorrhagic *E. coli* (EHEC) and is a significant food-borne pathogen worldwide.[6,7] Other important EHEC O serogroups include O26, O103, and O111.[8] The Shiga toxin-producing *E. coli* O104:H4 was found responsible for a large human food-borne disease outbreak in Europe, 2011.[9] Another notable example is strains of serogroup O25; extended-spectrum beta lactamase (ESBL)-producing, multidrug-resistant *E. coli* O25:H4 has emerged worldwide to cause a wide variety of community and nosocomial infections.[10]

In *E. coli*, the genes required for O-antigen biosynthesis are clustered at a chromosomal locus flanked by the colanic acid biosynthesis gene cluster (*wca* genes) and the histidine biosynthesis (*his*) operon. Generally, the O-antigen biosynthesis genes fall into three classes: (i) the nucleotide sugar biosynthesis genes, (ii) the sugar transferase genes, and (iii) those for O-unit translocation and chain synthesis (*wzx*/*wzy* in the Wzx/Wzy-dependent pathway and *wzm*/*wzt* in the Wzm/Wzt-dependent ABC transporter pathway).[11] To date, >90 types of O-antigen biosynthesis gene cluster (O-AGC) sequences have been determined, with the majority derived from major human and animal pathogens.[12] Sequence comparisons of these O-AGCs indicate a great variety of genetic structures. Several studies have provided evidence to show that horizontal transfer and replacement of a part or all of the O-AGC have caused shifts in O serogroups.[13–15] Alternatively, point mutations in the glycosyltransferase genes in the O-AGC or acquisition of alternative O-antigen modification genes, which are located outside of the O-AGC, have also been shown to result in structural alterations of O antigen and concomitant change in the serotype of the isolate.[16,17]

Genes or DNA sequences specific for each O serogroup can be used as targets for the identification of O serogroups via molecular approaches, such as PCR-based and hybridization-based methods. Such systems have already been developed by several researchers to target specific O-antigen types.[12,18–20] In particular, molecular assays targeting major O serogroups are routinely used in EHEC surveillance for clinical or food sample screening. Considering the range of diseases caused by *E. coli* strains belonging to many different serogroups, a more comprehensive and detailed O-AGC information for the complete set of *E. coli* O serogroups is of significant clinical importance for generating a rational molecular typing scheme. This molecular typing scheme, which could be performed *in silico* directly on sequence data, also offers a mechanism with which to link the ever-expanding genomic data to our extensive epidemiological and biological knowledge of this pathogen, based on O-antigen typing. Moreover, these data will also provide a much better understanding of the complex mechanisms by which a huge diversity in O serogroups have arisen. Here, we present a complete sequence set for the O-AGCs from all 184 *E. coli* O serogroups, which include recently added serogroups (O182–O187), providing a complete picture of the O-AGC diversity in *E. coli*.

2. Materials and methods

2.1. Bacterial strains, culture condition, and DNA preparation

Reference strains of all 184 recognized *E. coli* O serogroups were obtained from SSI (see Supplementary Table S1). Cells were grown to the stationary phase at 37°C in Luria–Bertani medium. Genomic DNA was purified using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions.

2.2. O-AGC sequences and comparative analyses

One hundred and eight *E. coli* O-AGC sequences were determined by Sanger-based capillary sequencing and/or Illumina MiSeq sequencing from PCR products covering O-AGCs (Supplementary Table S1). The O-AGC regions of the reference strains were amplified by PCR using 10 ng of genomic DNA as template with the Tks Gflex DNA polymerase (Takara Bio Inc.) by 25 amplification cycles for 10 s at 98°C and for 16 m at 69°C, and with a combination of three forward primers (TATGCCAGCGGCACCAAACG, ATACCGGCGATGAAAGCC, and GCGGGTGGGATTAAGTCTCT) designed on the *hisFI* genes and two reverse primers (GTGATGCAGGAATCCTCTGT and CCACGCTAATTACGCCATCTT) designed on the *wcaM* genes, or strain-specific primers designed based on the draft genome sequences determined using the MiSeq system from reference strains. Identification and functional annotation of the CDSs were performed based on the results of homology searches against the public, non-redundant protein database using BLASTP. The sequences reported in this article have been deposited in the GenBank database (accession no. AB811596-AB811624, AB812020-AB812085, and AB972413-AB972425). The other 76 *E. coli* O-AGC sequences were obtained from public databases. For a list of accession numbers, see Supplementary Table S1.

2.3. Phylogenetic analysis

Multilocus sequence typing (MLST) was carried out according to the protocol described on the *E. coli* MLST website (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli), and the phylogenetic relationships of reference strains were analysed based on the concatenated sequences (3,423 bp) of seven housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*) used for MLST. Multiple alignments of DNA and amino acid sequences were constructed by using the CLUSTAL W program.[21] Phylogenetic trees were constructed by using the neighbour-joining algorithm using the MEGA4 softrware.[22]

3. Results

3.1. Genetic structures of the O-AGCs from all *E. coli* O serogroups

Of the 184 known O serogroups, 76 complete O-AGC sequences were obtained from public databases. The sequence of the other 108 O-AGC was determined in this study from *E. coli* O-serogroup reference strains (Supplementary Table S1). Our analysis of these sequences confirmed several previously observed characteristics of O-AGCs in *E. coli* (Supplementary Fig. S1). In brief, O-AGCs are located between the *wca* and *his* operons. This region contains three housekeeping genes: *galF* (encoding UTP-glucose-1-phosphate uridylyltransferase), *gnd* (6-phosphogluconate dehydrogenase), and *ugd* (UDP-glucose 6-dehydrogenase), and most genes for O-antigen biosynthesis in each cluster are directly flanked by *galF* and *gnd*/*ugd*, while *gne* (UDP-GalNAc-4-epimerase) and *wzz* (O-antigen chain

length determination protein) located immediately outside of the region between *galF* and *gnd/ugd* (see Supplementary Fig. S1). The exceptions for this are the O-AGCs for O serogroups O14 and O57, which contain no O-antigen genes at the typical locus. However, it is known that the *E. coli* O14 reference strain Su4411-41 shows an O rough phenotype and lacks the O-AGC.[23] For O57, a further analysis is also required to investigate the presence of O-antigen structure in the LPS of the reference strain. Our data revealed that the O-AGCs located between *galF* and *gnd* ranged in size from 4.5 kbp (O155, including four genes) to 19.5 kbp (O108, including 18 genes).

3.1.1. Nucleotide sugar biosynthesis genes

Genes required for the deoxythymidine diphosphate (dTDP)-sugar biosynthesis pathway (*rmlBDAC*) to synthesize dTDP-L-rhamnose (dTDP-L-Rha), the precursor of L-Rha, were widely distributed in the O-AGCs (conserved in 56 O-serogroup O-AGCs; see Supplementary Fig. S1). The *vioAB* operon, for the biosynthesis of dTDP-*N*-acetylviosamine (dTDP-VioNAc), the precursor of VioNAc, was present in three O-serogroup O-AGCs; the *fnlABC* operon for the synthesis of uridine diphosphate (UDP)-*N*-acetyl-L-fucosamine (UDP-L-FucNAc), the precursor of L-FucNAc, was in 11 O-serogroup O-AGCs; the *fnlA-qnlBC* genes for the synthesis of UDP-*N*-acetyl-L-quinovosamine (UDP-L-QuiNAc), the precursor of L-QuiNAc, were in four O-serogroup O-AGCs; the *nnaDBCA* genes for synthesis of cytidine monophosphate (CMP)-*N*-acetylneuraminate (CMP–NeuNAc), the precursor of *N*-acetylneuraminic acid (Neu5Ac or sialic acid), were found in six O-serogroup O-AGCs (Supplementary Fig. S1). In addition, a gene set comprising seven genes putatively involved in the synthesis of di-*N*-acetyl-8-epilegionaminic acid (8eLeg5Ac7Ac) were found in three O-serogroup O-AGCs. For at least 49 O serogroups, gene sets for nucleotide sugar biosynthesis were not found in their O-AGCs (Supplementary Fig. S1), suggesting that, in these serogroups, nucleotide sugars required for O-antigen biosynthesis were synthesized by pathways encoded by the genes located outside of the O-AGCs.

3.1.2. Glycosyltransferase

Each O-AGC contained two to six genes encoding putative glycosyltransferases for synthesizing O-antigen subunits and a total of 611 glycosyltransferase genes identified in all O-AGCs. Pfam analysis revealed that at least 25 types of glycosyltransferase-related domains were found in the 611 glycosyltransferase genes (Supplementary Table S2). 'Glycosyl transferases group 1' (PF00534) and 'Glycosyl transferase family 2' (PF00535) were the most widely distributed domains, which were found in 216 and 253 genes, respectively. Except for the five genes belonging to 'Glycosyltransferase family 52' (PF07922), which were found in five of the six *nnaDBCA*-containing O-AGCs (O24, O56, O104, O131, and O171), there were no relationships between the type of glycosyltransferase-related domain and the gene set for sugar synthesis in each O-AGC.

3.1.3. O-antigen subunit translocation and chain synthesis

All O-AGCs carried either *wzx/wzy* or *wzm/wzt* gene pairs. Of the 182 O-AGCs (the above-mentioned O14 and O57 were excluded from the 184 clusters analysed in this study), 171 carried the *wzx/wzy* genes, and the other 11 carried the *wzm/wzt* genes (Supplementary Fig. S1 and Table S1). Detailed sequence comparisons of the *wzx/wzy* and *wzm/wzt* genes are described below.

3.2. Grouping the O-AGCs by sequence

On the basis of sequences and genetic structures of the entire O-AGC regions, in addition to 145 unique O-AGCs from different *E. coli* O serogroups, the O-AGCs from 37 O serogroups could be placed into 16 groups (named Gp1–Gp16) with the members of each group having identical or very similar O-AGC genes (mostly sharing ≥95% DNA sequence identity) (Fig. 1). This included nine groups with members of different serogroups but which carried identical O-AGC gene sets (Gp1–Gp9) and one group, Gp10, where two strains (O13 and O129) of the three-member group carried an identical O-AGC gene set (sharing 98.3–99.9% DNA sequence identity) (Fig. 1). The reason(s) why they belong to different O serogroups even though they have identical O-AGCs are discussed in the Discussion section. Indels or exchange of one or more genes was also shown to explain the differences between O135 and other members of Gp10 and members Gp11–Gp16, which otherwise carried highly conserved orthologous genes (summarized in Fig. 1). Simple insertions of insertion sequence (IS) elements containing one or two transposase genes were found in three groups without any gene disruption: an IS*629* insertion in O18ab of Gp12, IS*Ec11* in O164 of Gp13, and IS*1* in O62 of Gp14. IS element-associated replacement of the right-end portion of the O-AGC had occurred in three groups, Gp14, Gp15, and Gp16, resulting in the replacement (or deletion) of glycosyltransferase gene(s). Exchange of the *wzx* gene had also occurred in Gp16. These data suggest that IS elements are important drivers for generating O-antigen biosynthesis gene replacement and therefore diversity.

3.3. Diversity and specificity of the *wzx/wzy* or *wzm/wzt* genes among the *E. coli* O-AGCs

As previously proposed,[12] most *wzx/wzy* or *wzm/wzt* orthologues showed high levels of sequence diversity and their sequences were unique to each O-AGC or O-AGC group described above (Fig. 2 and Supplementary Fig. S2). DNA sequence identities of the closest pairs were <70%, except for the O96/O170 pair, the *wzx* genes of which showed 86% DNA sequence identity. Within the 16 O-AGC groups, the orthologous *wzx/wzy* or *wzm/wzt* genes also showed high sequence conservation (≥95% DNA sequence identity, but mostly ≥97% identity), except for Gp16 that shared only the *wzy* gene (Fig. 2).

3.4. Phylogenetic relationships of *E. coli* O-serogroup reference strains

Based on the concatenated nucleotide sequences of seven housekeeping genes used for MLST, we determined the evolutionary relationships of all *E. coli* O-serogroup reference strains (Fig. 3). This analysis revealed that the members of five groups sharing the common O-AGCs (Gp8, Gp10, Gp11, Gp14, and Gp15) and two members (O17 and O77) of Gp9 were found in closely related lineages. However, the members of other groups (and three members of Gp9) were found in distinct evolutionary lineages. For example, O20 and O137, both carrying the Gp1 O-AGC, were found in two distinct lineages, each belonging to phylogroups A and E/D, respectively, and five serogroups (O17/O77, O44, O73, and O106) belonging to Gp9 were found in multiple lineages (A, E/D, and B1).

The systematic phylogenetic analysis of all *E. coli* O-serogroup reference strains further revealed that one-quarter of the reference strains (46/184) belonged to a single clonal group (≥99.9% sequence identity), which was represented by sequence type (ST) 10 and its very close relatives in phylogroup A (Fig. 3 and Supplementary Fig. S3). Additionally, three clonal groups containing five or more reference

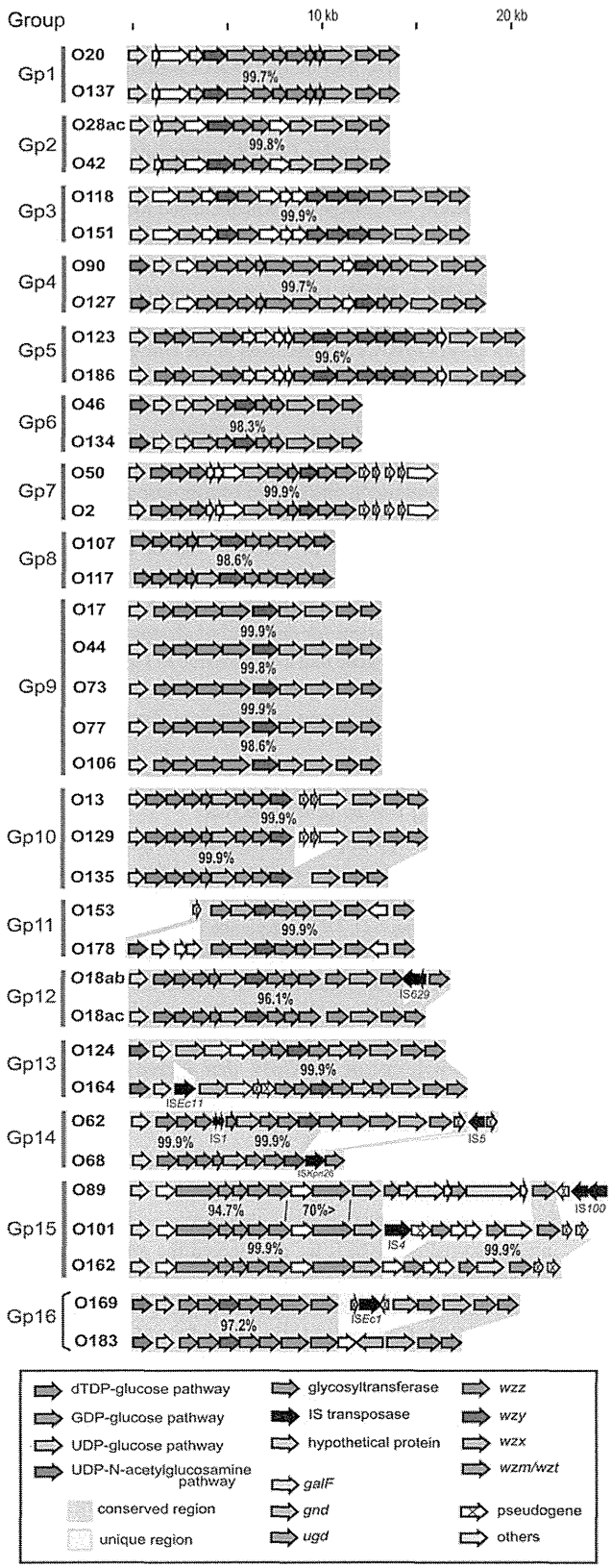

Figure 1. Sixteen *Escherichia coli* O-AGC groups identified in this study. Group members have different O serogroups in each group, but these share nearly identical or highly similar genetic organizations. Group names (Gp) are indicated at the left side. DNA sequence identities (%) between group members are indicated in each group.

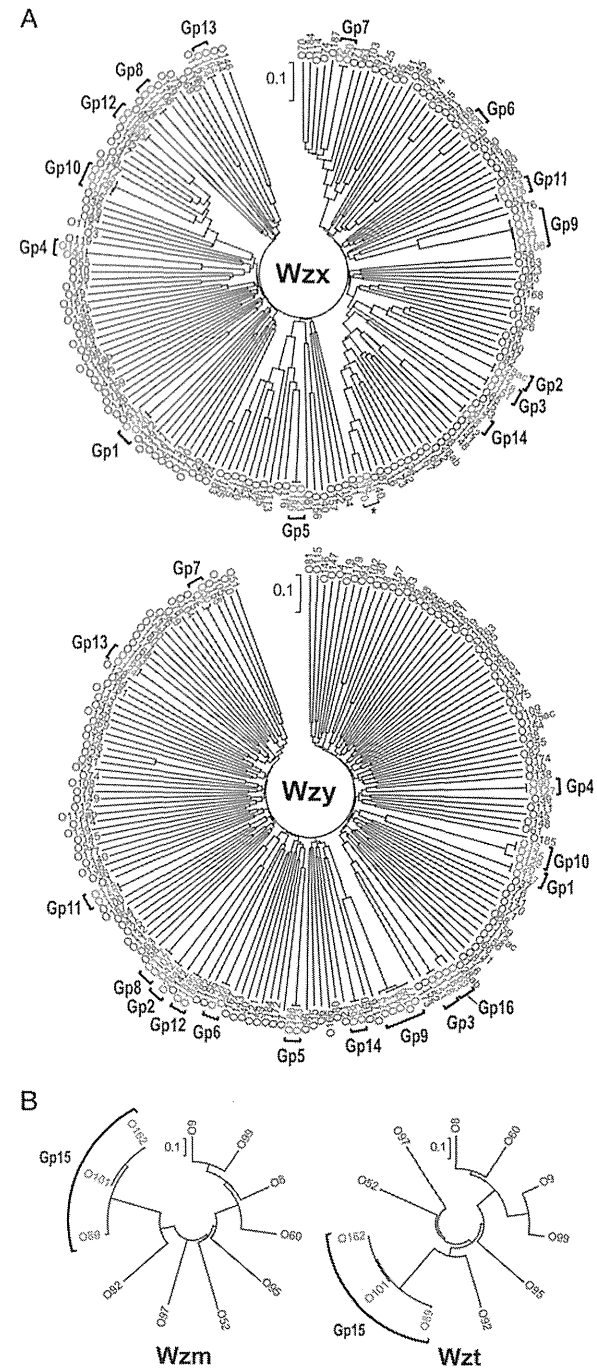

Figure 2. Phylogenetic analysis of homologues of (A) Wzx and Wzy and (B) Wzm and Wzt from *Escherichia coli* O-serogroup reference strains based on the amino acid sequences. The group names are indicated outside of trees. The pair or groups of homologues with high DNA sequence identity (≥95%, mostly ≥97%) are indicated in red. The Wzx homologues of O96 and O170, which are indicated in blue and by an asterisk, showed 86% DNA sequence identity, but in all other proteins showed low-sequence homologies to each other (<70% identity). Note that while the DNA sequence identity between the *wzx*_O46 and *wzx*_O134 in Gp6 is 99.7%, the *wzx*_O46 has a 2-bp deletion at the 3′-region, causing a frame shift.

strains were also identified in phylogroups A (ST34 and ST57) and B1 (ST300) (Fig. 3). The phylogenetic analysis also showed that the types of sugar synthesis gene sets and processing gene sets (*wzx/wzy* and *wzm/wzt*) were not limited to a specific lineage (Fig. 3).

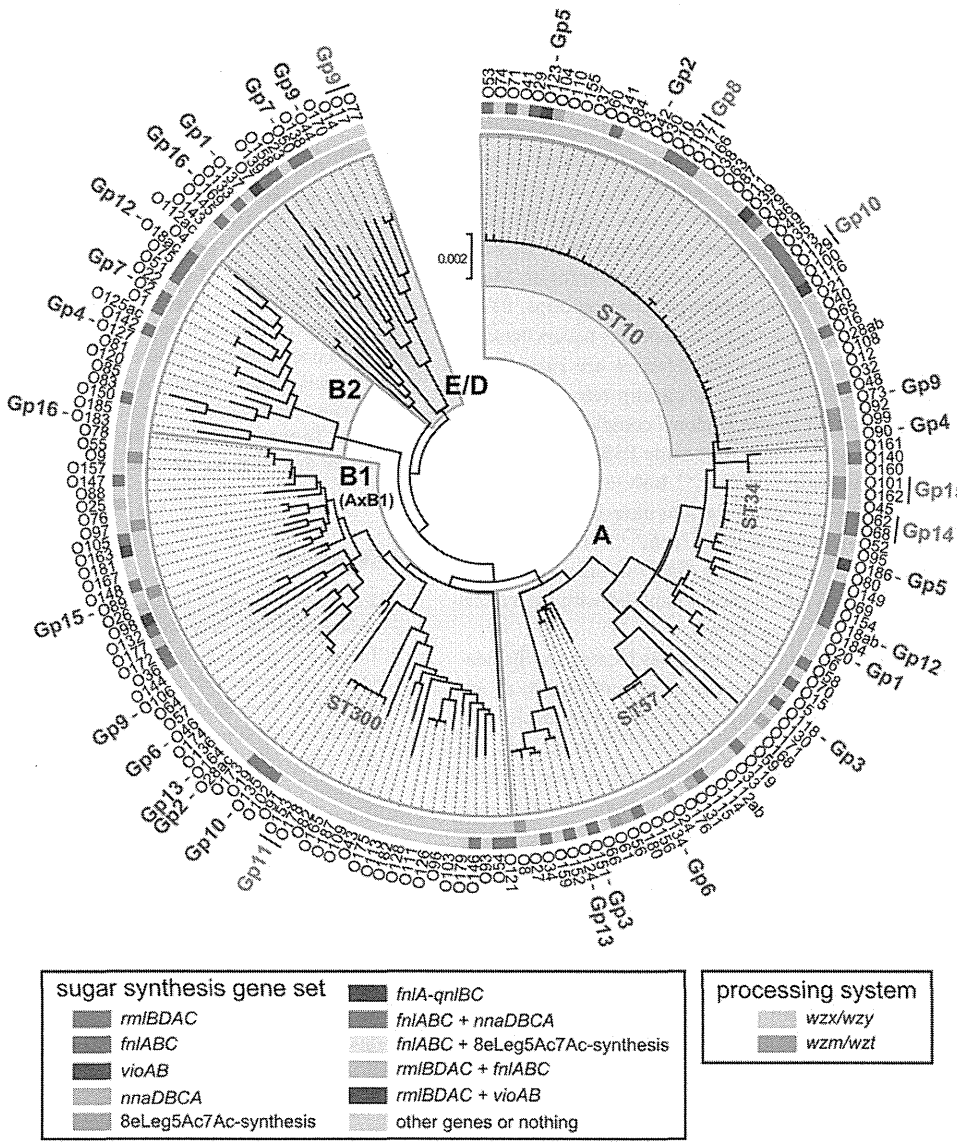

Figure 3. Correlation between the *Escherichia coli* evolutionary lineages and the distribution of O-AGCs. The phylogenetic tree was constructed based on the concatenated sequences of seven housekeeping genes from all 184 *E. coli* O-serogroup reference strains. The group names of O-AGCs (Gp1–Gp16) are indicated in the outermost region. Members in groups indicated in green were found to belong to the same or very closely related lineage, whereas members of the groups indicated in blue were found in distinct lineages. The outer circle next to the O serogroup names indicates the distribution of sugar synthesis gene sets identified in each O-AGC. The inner circle indicates the type of O-antigen processing system (*wzx/wzy* or *wzm/wzt*). Phylogenetic groups (A, B1, B2, D, and E) were determined by comparing the sequences of the strains tested with the known sequences from the ECOR collection (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli).

3.5. Relationships of the *E. coli* and *Shigella* O-AGCs

Shigella and *E. coli* belong to the same species complex[24] and many *Shigella* O antigens are known to be serologically and genetically identical or very similar to some *E. coli* O antigens, as summarized by Liu *et al.*[25] In addition to the 21 previously shown relationships, we found two additional O-AGC groups shared by *E. coli* and *Shigella*; O38 and *Shigella dysenteria* type 8 (SD8), and O169/O183 and *Shigella boydii* type 6/10 (SB6/SB10) (Supplementary Fig. S4). The O183-AGC was highly similar to the *S. boydii* types 10 cluster (sharing 98.2% DNA sequence identity). In our previous study,[26] we provisionally named a novel O serogroup for a group of Shiga toxin-producing *E. coli* strains as OSB10, which cross-reacted with *S. boydii* type 10. Sequence comparisons in this study revealed that OSB10 is not only serologically but also genetically identical to the new serogroup O183 of Gp16.

4. Discussion

Much of what we know about *E. coli* is defined at some level by O serogroups. To link genomic information to the wealth of data held in public databases, in our collective knowledge, outbreak, and disease reports and elsewhere, we endeavoured to determine whether molecular O-serogroup identification, targeting O-serogroup-specific genes (or unique sequences), was a valuable method to capture this information and maintain this important link. Not only do we show evidence supporting the effectiveness of molecular O-typing, but also we open

up the possibility of generating a molecular O-typing scheme and relate O serogroups to the underlying phylogeny of this bacterium.

By determining and comparing the sequences of O-AGCs from all known *E. coli* O serogroups, we newly defined the sequence and gene content of 145 unique O-AGCs and showed that O-AGCs from 37 O serogroups could be placed into 16 groups based on members in each group sharing nearly identical or highly similar O-AGCs. It is clear from these data that many of the grouped O-AGCs (Gp1-16) were found in distinct phylogenetic lineages indicating that these O-AGCs have been spread across this species by horizontal gene transfer. Moreover, several lineages that contained multiple O serogroups, ST10, ST34, ST57, and ST300, show that frequent exchange occurs between and within lineages. ST10 and its close relatives are particularly interesting as one-quarter of *E. coli* O-serogroup reference strains fell within this clonal group. ST10 and its clonal complex are clinically very important being recently found to include ESBL-producing *E. coli* from human and animals in Spain,[25] Italy and Denmark,[26] China,[27] and the Netherlands,[28] and in various intra-intestinal pathotypes of *E. coli*, such as enteroaggregative *E. coli*,[27,28] enterotoxigenic *E. coli*,[29,30] and EHEC.[31,32] In most cases, the O serogroups of these ST10 or ST10-related strains are unusual compared with the typical O serogroups that represent that pathotype.

Acquisition of O-antigen modification genes located on the genomes of serotype-converting bacteriophages or plasmids is also an important strategy for diversifying O-antigen structures. This mechanism has been well investigated in *Shigella flexneri*.[33,34] In *E. coli*, the O-serogroup conversion by a prophage-like element has been reported for O17 and O44,[17] which belong to Gp9 defined in this study. Another possible mechanism to generate the variation of O antigens is the mutations in the genes of the O-AGC as observed for O107 and O117,[16] which belong to Gp8. In this case, point mutations in a glycosyltransferase gene are responsible for the alteration of O-antigen structure (and thus that of O serogroup).[16] Five O-AGC groups including Gp2, Gp5, Gp7, Gp12, and Gp13 also contained differences in the amino acid sequence of their glycosyltransferases. O serogroup differences in these groups may be generated by the point mutations in glycosyltransferase genes. On the other hand, all glycosyltransferase genes in Gp1, Gp3, Gp4, Gp6, and Gp11; four strains from Gp9 (O17, O44, O73, and O77) and two from Gp10 (O13 and O129) showed 100% amino acid sequence identity. These results suggest that the serological differences between the members of these seven groups have been generated by acquisition of modification genes outside of the O-AGC as shown for O17 and O44 of Gp9.[17]

We believe that the remarkable sequence diversity observed in the *wzx/wzy* and *wzm/wzt* O-AGC genes of all known *E. coli* O serogroups appears to be sufficiently discriminative from one another to make identification of each of the known O serogroups possible. Therefore, our sequence data will serve as a valuable resource for the development of rationally designed molecular methods for O-typing as well as for detecting novel O serogroups.

In conclusion, our study provides a complete sequence set of O-AGCs of all known *E. coli* O serogroups and thus offers a full view on the genetic diversity of O-AGCs of this bacterium. In addition, the results presented suggest that horizontal gene transfer has been involved in the O serogroup diversification in *E. coli* more frequently and in a more biased or lineage-dependent fashion than previously thought.

Acknowledgements

We thank A. Akiyoshi, Y. Kato, and A. Yoshida for technical assistance.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by Health Labor Sciences Research Grants from the Ministry of Health, Labor, and Welfare, Japan to A.I. (H25-Syokuhin-Wakate-018) and M.O. (H24-Shinkou-Ippan-012); Adaptable and Seamless Technology Transfer Program through Target-driven R&D (AS242Z00217P) from Japan Science and Technology Agency to A.I.; and a Scientific Research Grant on Priority Areas from the University of Miyazaki and the Program to Disseminate Tenure Tracking System from the Japanese Ministry of Education, Culture, Sports, Science, and Technology to A.I. (http://www.miyazaki-u.ac.jp/ir/english/index.html). This work was also supported by Wellcome Trust grant (098051). Funding to pay the Open Access publication charges for this article was provided by the University of Miyazaki, Japan.

References

1. Bazaka, K., Crawford, R.J., Nazarenko, E.L. and Ivanova, E.P. 2011, Bacterial extracellular polysaccharides, *Adv. Exp. Med. Biol.*, 715, 213–26.
2. Liu, B., Knirel, Y.A., Feng, L., et al. 2013, Structural diversity in *Salmonella* O antigens and its genetic basis, *FEMS Microbiol. Rev.*, 38, 56–89.
3. Stenutz, R., Weintraub, A. and Widmalm, G. 2006, The structures of *Escherichia coli* O-polysaccharide antigens, *FEMS Microbiol. Rev.*, 30, 382–403.
4. Lam, J.S., Taylor, V.L., Islam, S.T., Hao, Y. and Kocincova, D. 2011, Genetic and functional diversity of *Pseudomonas aeruginosa* lipopolysaccharide, *Front Microbiol.*, 2, 118.
5. Penner, J.L. and Aspinall, G.O. 1997, Diversity of lipopolysaccharide structures in *Campylobacter jejuni*, *J. Infect. Dis.*, 176 (Suppl. 2), S135–138.
6. Armstrong, G.L., Hollingsworth, J. and Morris, J.G. Jr. 1996, Emerging foodborne pathogens: *Escherichia coli* O157:H7 as a model of entry of a new pathogen into the food supply of the developed world, *Epidemiol Rev.*, 18, 29–51.
7. Tarr, P.I., Gordon, C.A. and Chandler, W.L. 2005, Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome, *Lancet*, 365, 1073–86.
8. Johnson, K.E., Thorpe, C.M. and Sears, C.L. 2006, The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*, *Clin. Infect. Dis.*, 43, 1587–95.
9. Buchholz, U., Bernard, H., Werber, D., et al. 2011, German outbreak of *Escherichia coli* O104:H4 associated with sprouts, *N. Engl. J. Med.*, 365, 1763–70.
10. Peirano, G. and Pitout, J.D. 2010, Molecular epidemiology of *Escherichia coli* producing CTX-M beta-lactamases: the worldwide emergence of clone ST131 O25:H4, *Int. J. Antimicrob. Agents*, 35, 316–21.
11. Samuel, G. and Reeves, P. 2003, Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly, *Carbohydr. Res.*, 338, 2503–19.
12. DebRoy, C., Roberts, E. and Fratamico, P.M. 2011, Detection of O antigens in *Escherichia coli*, *Anim. Health Res. Rev.*, 12, 169–85.
13. Leopold, S.R., Magrini, V., Holt, N.J., et al. 2009, A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis, *Proc. Natl Acad. Sci. USA*, 106, 8713–8.
14. Iguchi, A., Shirai, H., Seto, K., et al. 2011, Wide distribution of O157-antigen biosynthesis gene clusters in *Escherichia coli*, *PLoS ONE*, 6, e23250.
15. Iguchi, A., Iyoda, S. and Ohnishi, M. 2012, Molecular characterization reveals three distinct clonal groups among clinical Shiga toxin-producing *Escherichia coli* strains of serogroup O103, *J. Clin. Microbiol.*, 50, 2894–900.
16. Wang, Q., Perepelov, A.V., Wen, L., et al. 2012, Identification of the two glycosyltransferase genes responsible for the difference between *Escherichia coli* O107 and O117 O-antigens, *Glycobiology*, 22, 281–7.

17. Wang, W., Perepelov, A.V., Feng, L., et al. 2007, A group of *Escherichia coli* and *Salmonella* enterica O antigens sharing a common backbone structure, *Microbiology*, **153**, 2159–67.
18. Lacher, D.W., Gangiredla, J., Jackson, S.A., Elkins, C.A. and Feng, P.C. 2014, Novel microarray design for molecular serotyping of Shiga toxin-producing *Escherichia coli* isolated from fresh produce, *Appl. Environ. Microbiol.*, **80**, 4677–82.
19. Tzschoppe, M., Martin, A. and Beutin, L. 2012, A rapid procedure for the detection and isolation of enterohaemorrhagic *Escherichia coli* (EHEC) serogroup O26, O103, O111, O118, O121, O145 and O157 strains and the aggregative EHEC O104:H4 strain from ready-to-eat vegetables, *Int. J. Food Microbiol.*, **152**, 19–30.
20. Wang, Q., Ruan, X., Wei, D., et al. 2010, Development of a serogroup-specific multiplex PCR assay to detect a set of *Escherichia coli* serogroups based on the identification of their O-antigen gene clusters, *Mol. Cell Probes*, **24**, 286–90.
21. Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, **22**, 4673–80.
22. Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007, MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, *Mol. Biol. Evol.*, **24**, 1596–9.
23. Jensen, S.O. and Reeves, P.R. 2004, Deletion of the *Escherichia coli* O14:K7 O antigen gene cluster, *Can. J. Microbiol.*, **50**, 299–302.
24. Pupo, G.M., Lan, R. and Reeves, P.R. 2000, Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics, *Proc. Natl Acad. Sci. USA*, **97**, 10567–72.
25. Liu, B., Knirel, Y.A., Feng, L., et al. 2008, Structure and genetics of *Shigella* O antigens, *FEMS Microbiol. Rev.*, **32**, 627–53.
26. Iguchi, A., Iyoda, S., Seto, K. and Ohnishi, M. 2011, Emergence of a novel Shiga toxin-producing *Escherichia coli* O serogroup cross-reacting with *Shigella boydii* type 10, *J. Clin. Microbiol.*, **49**, 3678–80.
27. Olesen, B., Scheutz, F., Andersen, R.L., et al. 2012, Enteroaggregative *Escherichia coli* O78:H10, the cause of an outbreak of urinary tract infection, *J. Clin. Microbiol.*, **50**, 3703–11.
28. Okeke, I.N., Wallace-Gadsden, F., Simons, H.R., et al. 2010, Multi-locus sequence typing of enteroaggregative *Escherichia coli* isolates from Nigerian children uncovers multiple lineages, *PLoS ONE*, **5**, e14093.
29. Turner, S.M., Chaudhuri, R.R., Jiang, Z.D., et al. 2006, Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages, *J. Clin. Microbiol.*, **44**, 4528–36.
30. Nada, R.A., Shaheen, H.I., Khalil, S.B., et al. 2011, Discovery and phylogenetic analysis of novel members of class b enterotoxigenic *Escherichia coli* adhesive fimbriae, *J. Clin. Microbiol.*, **49**, 1403–10.
31. Monaghan, A.M., Byrne, B., McDowell, D., Carroll, A.M., McNamara, E. B. and Bolton, D.J. 2012, Characterization of farm, food, and clinical Shiga toxin-producing *Escherichia coli* (STEC) O113, *Foodborne Pathog. Dis.*, **9**, 1088–96.
32. Hauser, E., Mellmann, A., Semmler, T., et al. 2013, Phylogenetic and molecular analysis of food-borne shiga toxin-producing *Escherichia coli*, *Appl. Environ. Microbiol.*, **79**, 2731–40.
33. Allison, G.E. and Verma, N.K. 2000, Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*, *Trends Microbiol.*, **8**, 17–23.
34. Sun, Q., Knirel, Y.A., Lan, R., et al. 2012, A novel plasmid-encoded serotype conversion mechanism through addition of phosphoethanolamine to the O-antigen of *Shigella flexneri*, *PLoS ONE*, **7**, e46095.

Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution

Astrid von Mentzer[1,2], Thomas R Connor[2,3], Lothar H Wieler[4], Torsten Semmler[4], Atsushi Iguchi[5], Nicholas R Thomson[2], David A Rasko[6], Enrique Joffre[1], Jukka Corander[7], Derek Pickard[2], Gudrun Wiklund[1], Ann-Mari Svennerholm[1], Åsa Sjöling[1,8] & Gordon Dougan[2]

© 2014 Nature America, Inc. All rights reserved.

Enterotoxigenic *Escherichia coli* (ETEC), a major cause of infectious diarrhea, produce heat-stable and/or heat-labile enterotoxins and at least 25 different colonization factors that target the intestinal mucosa. The genes encoding the enterotoxins and most of the colonization factors are located on plasmids found across diverse *E. coli* serogroups. Whole-genome sequencing of a representative collection of ETEC isolated between 1980 and 2011 identified globally distributed lineages characterized by distinct colonization factor and enterotoxin profiles. Contrary to current notions, these relatively recently emerged lineages might harbor chromosome and plasmid combinations that optimize fitness and transmissibility. These data have implications for understanding, tracking and possibly preventing ETEC disease.

ETEC cause approximately 400 million diarrheal cases and almost 400,000 deaths per year in children aged less than 5 years in low- and middle-income countries and are also a common cause of travelers' diarrhea[1]. ETEC are defined by their ability to produce a heat-labile toxin (LT) and/or a heat-stable toxin (ST; including two subtypes, STh and STp)[2,3]. At least 25 antigenically distinct colonization factors have been described in human ETEC. Colonization factors are fimbrial or afimbrial surface structures with the potential to mediate adherence to the human intestinal mucosa[2]. The most prevalent colonization factors are CFA/I and coli surface antigens 1–6 (CS1–CS6), although in certain geographical regions CS7, CS14 and CS17 are also common[2]. Individual ETEC isolates typically carry and/or coexpress one, two or three colonization factors and/or toxin types, with combinations such as CS1 + CS3 with LT + STh, CS2 + CS3 with LT + STh, CS5 + CS6 with LT + STh, CS6 with STp, CFA/I with STh and CS7 with LT repeatedly isolated globally[2–5]. However, 20–50% of all clinical ETEC isolates, in particular, those expressing LT only, do not express any of the identified colonization factors, suggesting that additional colonization factors might exist[4,5].

In addition to the large number of identified colonization factors and colonization factor–toxin combinations, ETEC can express a wide variety of O antigens (over 100 different O antigens have been associated with clinical ETEC isolates)[2,3,6]. Together with the large number of colonization factors, this wide range of O antigens indicates that there is a substantial level of genetic diversity within this pathovar[6]. Limited sequence-based studies of ETEC phylogeny have indicated that the acquisition of colonization factor and toxin genes by non-pathogenic, commensal strains might be sufficient to cause clinical ETEC disease[7,8]. Previous studies, exploring a potential association between chromosomal backgrounds and virulence factors in ETEC, have not shown any consistent evidence of phylogenetic clustering of isolates, although a potential association between virulence profiles and genetic backgrounds was suggested in some studies[8–14]. In addition, it has been concluded that the acquisition of virulence-related genes has occurred multiple times, consistent with the key virulence genes being encoded on mobile plasmids[8,10]. These observations have led to the hypothesis that ETEC are simply any *E. coli* lineage that can acquire, express and retain plasmids harboring colonization factors and/or toxins.

In this study, we have used next-generation sequencing to develop a more complete understanding of ETEC phylogeny and evolution. To this end, we have examined a global collection of ETEC isolated from 20 different countries in Asia, Africa, and North, Central and South America between 1980 and 2011 by whole-genome sequencing, serotyping and phylogenetic analyses. To cover the breadth of ETEC diversity, we selected ETEC isolates on the basis of their colonization factor and toxin profiles, including isolates lacking known colonization factors, as well as isolates from individuals from different age groups in endemic countries and travelers. Our analyses show that ETEC are widely distributed across the *E. coli* species. There is also a clear signature of several globally distributed ETEC lineages that show consistent long-term association with a specific O antigen and virulence gene repertoire.

[1]Department of Microbiology and Immunology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden. [2]Wellcome Trust Sanger Institute, Hinxton, UK. [3]Organisms and Environment Division, Cardiff University School of Biosciences, Cardiff University, Cardiff, UK. [4]Centre of Infection Medicine, Institute of Microbiology and Epizootics, Freie Universität Berlin, Berlin, Germany. [5]Interdisciplinary Research Organization, University of Miyazaki, Miyazaki, Japan. [6]Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [7]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. [8]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to A.v.M. (astrid.von.mentzer@gu.se), Å.S. (asa.sjoling@ki.se), A.-M.S. (ann-mari.svennerholm@microbio.gu.se) or G.D. (gd1@sanger.ac.uk).

Received 12 February; accepted 17 October; published online 10 November 2014; doi:10.1038/ng.3145

RESULTS

ETEC are distributed throughout the *E. coli* species

Considering the diversity of *E. coli* and to provide an accurate genome-wide phylogeny, we identified 1,429 genes representing the 'maximum common genome' (MCG) from the genomes of 47 *E. coli* isolates representing the known species diversity (Online Methods), with these genes also found in all ETEC strains sequenced in this study.

The ETEC phylogeny was constructed from the MCG alignments of 362 selected ETEC isolates with representative virulence profiles isolated from indigenous populations and travelers between 1980 and 2011 from different countries in Asia, Africa, and North, Central and South America. Also included were 21 available reference genomes covering commensal *E. coli*, enteropathogenic *E. coli* (EPEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), enterohemorrhagic *E. coli* (EHEC), uropathogenic *E. coli* (UPEC), *Shigella* and previously published ETEC genomes (**Fig. 1** and **Supplementary Tables 1** and **2**). An alignment was generated covering the 1,429 genes for the 383 genomes, which was then used as a basis for the detection of recombination sites, SNP calling and phylogenetic tree construction. In total, we identified 128,214 variable sites (excluding the reference genomes) across the MCG showing the diversity captured within this pathovar. This analysis demonstrated that ETEC are clearly distributed throughout the species phylogeny, consistent with previous studies[8] (**Fig. 1**). Indeed, ETEC isolates were assigned to most recognized phylogroups of *E. coli*[15], with the majority falling within the A and B1 groups (**Fig. 1**).

Identification of several major ETEC lineages

The MCG-based phylogenetic analysis together with Bayesian analysis of the population structure (BAPS)[16] was employed to define lineages across ETEC. The BAPS analysis defined several robust ETEC lineages (L1–L21) among the 362 ETEC isolates analyzed (**Fig. 1**). The majority of the lineages encompassed isolates obtained from various countries in Asia, the Americas and Africa collected over three decades.

Lineages share specific virulence factors and are globally spread

The colonization factor and toxin profiles of all 362 ETEC isolates sequenced were mapped onto the MCG tree, and the presence of colonization factors and toxins was reconfirmed using comparative genomics approaches (Online Methods). Known colonization factors were identified in 228 ETEC isolates, whereas 134 isolates lacking an identifiable colonization factor (by phenotypic and genomic

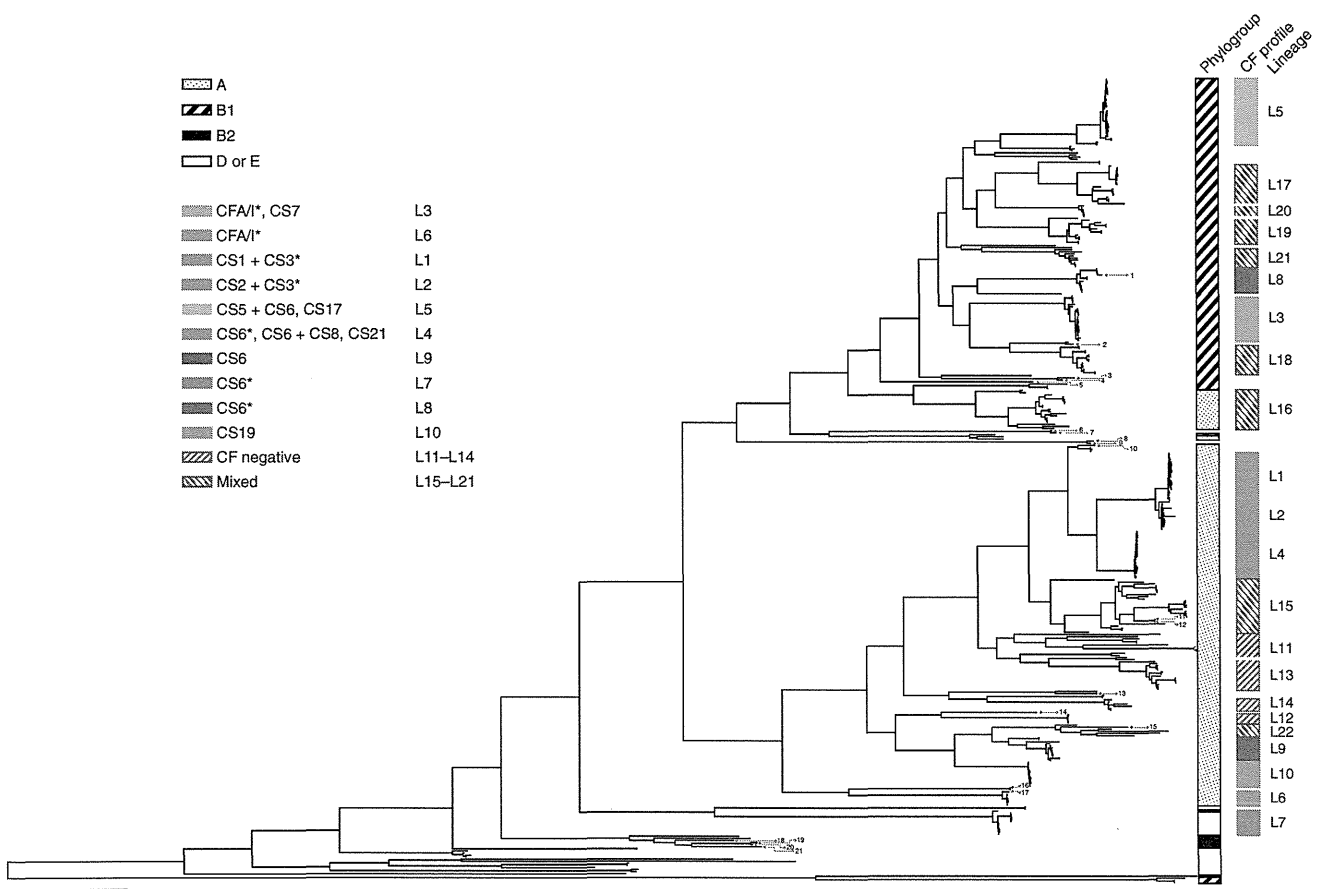

Figure 1 Population structure of ETEC isolates. Midpoint-rooted maximum-likelihood phylogenetic tree based on SNP differences across the MCG, excluding probable recombination events. The phylogenetic groups of isolates are shown in black and white to the right of the tree. Distinct lineages identified with BAPS across the data set are indicated to the right of the color-coded colonization factor profile, which match the lineages identified. An asterisk indicates isolates with or without CS21. References are indicated with red dots and arrows: 1, ETEC B7A; 2, ETEC 24377A; 3, EPEC B171; 4, EPEC E22; 5, EPEC E110019; 6, *S. sonnei* 53G; 7, *S. sonnei* Ss046; 8, *Shigella flexneri* 8401; 9, *S. flexneri* M90T; 10, ETEC H10407; 11, *E. coli* K-12 MG1655; 12, *E. coli* K12 W3110; 13, EAEC 101-1; 14, EIEC 53638; 15, *E. coli* HS; 16, EHEC EDL933; 17, EHEC Sakai; 18, UPEC CFT073; 19, UPEC 536; 20, UPEC F11; 21, UPEC UTI89. CF, colonization factor. Scale bar, 0.041 substitutions per variable site.

© 2014 Nature America, Inc. All rights reserved.

Table 1 Characteristics of ETEC isolates in lineages L1–L14

Lineage	Number of isolates[a]	Variable sites[b]	MLST[c]	O antigen	Colonization factor	Enterotoxin
L1	23	542	ST2353, ST4 ($n = 4$)	O6	CS1 + CS3[e]	LT + STh
L2	14	427	ST4	O6	CS2 + CS3[e]	LT + STh
L3	22	880	ST173	O78, O114, O126, O128	CFA/I[e], CS7	LT + STh, STh
L4	23	340	ST1312	O25	CS6[e], CS6 + CS8, CS21	LT, STh
L5	30	517	ST443	O115, O157	CS5 + CS6, CS17	LT + STh, LT, STh
L6	7	172	ST2332	ON3[d]	CFA/I[e]	STh
L7	12	483	ST182	O169	CS6	STp
L8	11	522	ST94	O148	CS6	STh, STp
L9	12	2,758	ST398	O27	CS6	STp
L10	12	57	ST2368	O114	CS19	LT + STp
L11	13	13,059	$n = 8$[f]	$n = 10$[f]	–	LT + STp, LT, ST
L12	5	80	ST731	O15	–	LT
L13	14	2,408	ST10, ST100, ST165, ST750, ST3860	ON49[d], O112ab, O160, O170, O179	–	LT + STp, LT
L14	6	502	ST10, ST1684, ST2705	ON5[d]	–	ST

[a]Number of isolates in L1–L5 used for Bayesian phylogenetic analysis with BEAST and in lineages L6–L14. [b]Determined on the basis of sequences with recombination sites removed. [c]MLST was determined by extracting the sequences for the *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA* and *recA* genes from the whole-genome data. [d]N, novel O antigen. [e]With or without CS21 (Longus). [f]Details on MLST and O antigens can be found in **Supplementary Table 2**.

© 2014 Nature America, Inc. All rights reserved.

analyses) were defined as 'colonization factor–negative' isolates. We also determined O antigen genotypes as predicted from sequence (**Fig. 1**, **Table 1** and Online Methods).

Interestingly, the isolates found in the major lineages L1–L5 also expressed the most prevalent virulence profiles described in the literature[5]. These lineages were selected for further analyses, as they comprised a large number of isolates appropriate for the subsequent studies. Lineages L1–L5 all showed a clear clustering of isolates with a specific virulence profile, i.e., a combined colonization factor and toxin profile. This clustering was also evident in several additional lineages (L6–L10) with distinct colonization factor and toxin profiles. We identified 38 colonization factor–negative isolates in 4 lineages (L11–L14). These observations indicate that ETEC is far more than a plasmid with an *E. coli* attached. The data also demonstrate that some ETEC isolates fall into distinct globally and temporally distributed lineages with specific virulence profiles (**Fig. 2**, **Table 1**, **Supplementary Figs. 1–4** and **Supplementary Table 2**). However, seven lineages (L15–L21) comprised ETEC isolates with a mix of colonization factor and toxin profiles, suggesting that in these lineages gene exchange might be common. All of the distinct ETEC lineages were represented by isolates taken from adults and children in endemic areas as well as from travelers with diarrhea (**Supplementary Table 2**).

The L1 lineage comprised ETEC with the colonization factor profile CS1 + CS3 (± CS21) (**Fig. 2** and **Table 1**). The closely related L2 lineage encompassed isolates expressing CS2 + CS3 (± CS21) and shared the O6 antigen with L1. Notably, isolates positive for CS1 + CS3 (± CS21) and CS2 + CS3 (± CS21) were not found in any other lineage (**Fig. 2** and **Table 1**). CFA/I-positive isolates were identified in two individual lineages, L3 and L6. Additional lineages, comprising isolates that shared O antigen and colonization factor and toxin profiles, were identified (**Table 1**). ETEC isolates expressing CS6 were found in four lineages (L4 and L7–L9). Ten CS19-positive isolates clustered together (L10), although these isolates were probably from a single geographically restricted outbreak.

There was also an interesting association of variation in O antigen genotype and colonization factor and toxin profile with phylogenetic lineage. On the basis of O antigen genotype, lineage L3 could be divided into four subclades: isolates encoding CFA/I with O78, O126 or O128, and five CS7 + O114 isolates. However, these isolates were related both in terms of colonization factor profile and phylogenetic origin. In contrast, isolates within the L4 lineage expressed CS6 alone or together with CS8 as well as CS21, but they all belonged to serogroup O25 (**Fig. 2** and **Table 1**). The largest ETEC lineage identified in this study was L5, which could be divided into two subclades on the basis of O antigen type. One subclade harbored O115-positive isolates that expressed either CS5 + CS6 or the distantly related CS17, whereas the second subclade harbored isolates positive for O167 and CS5 + CS6.

Colonization factor–negative isolates allocate across the ETEC tree

Four lineages (L11–L14) harboring mainly colonization factor–negative isolates were identified. In comparison to other lineages (L1–L10), which showed a substantial association between their O antigen and colonization factor and toxin profiles, the predominantly colonization factor–negative lineages showed a higher level of diversity. However, there was still structure in their phylogeny, suggesting that so-called colonization factor–negative isolates might share properties, including, for example, unidentified new colonization factors (**Fig. 1**, **Table 1**, **Supplementary Fig. 3** and **Supplementary Table 2**). Additional colonization factor–negative isolates were found across the tree, clustering together with isolates harboring less prevalent colonization factors such as CS12, CS14 and CS17 and, in some cases, CS6. The isolates in lineages L15–L21 represented 28% of our ETEC collection (**Fig. 1**, **Supplementary Fig. 3** and **Supplementary Table 2**). Hence, there are lineages with varied virulence profiles that have added to the confusion about this pathovar, but our data identify a number of persistent plasmid-chromosomal background combinations in ETEC, even in isolates without known colonization factors.

Toxin allele profiles are associated with chromosomal background

To further investigate a potential relationship between the chromosomal background, colonization factors and enterotoxins, we extracted the nucleotide sequences of the LT (*eltAB*) and ST (STh and STp genes) operons for further analyses. Sequence analysis showed that *eltAB* was more variable than the STh and STp genes, a finding in agreement with previous studies[17,18]. Mapping LT and ST allele data onto the MCG-based phylogenetic tree showed a close association of toxin alleles with the chromosomal background and the colonization factor and toxin profiles. For example, the closely related lineages L1

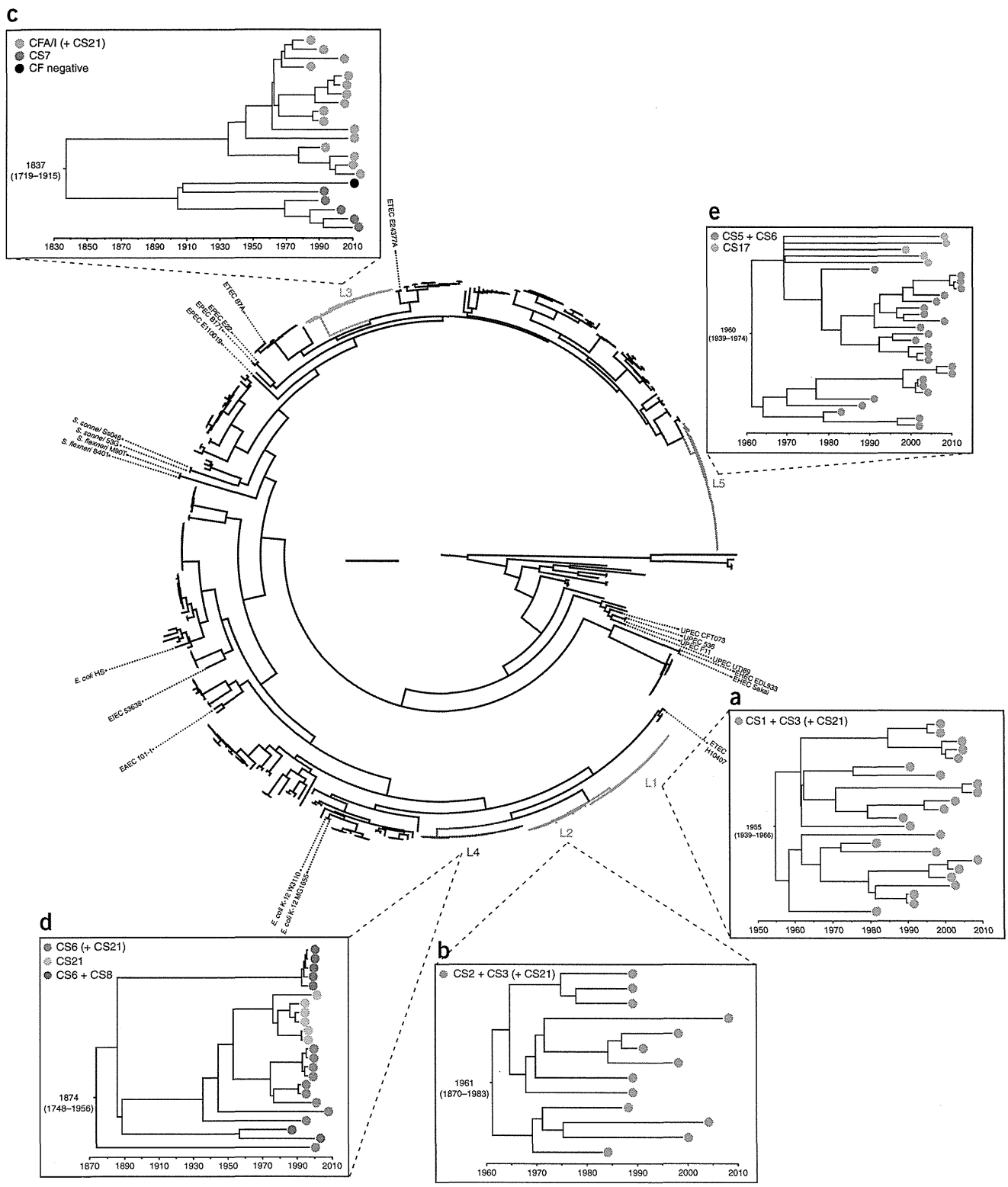

Figure 2 Most common recent ancestors of the five major lineages L1–L5. Midpoint-rooted circular phylogenetic tree of ETEC based on SNP differences, with probable recombination events removed and 21 *E. coli* references indicated. (**a**–**e**) Surrounding the circular tree are five individual SNP-based phylogenetic trees, with recombination events removed, of the major lineages: L1 (**a**), L2 (**b**), L3 (**c**), L4 (**d**) and L5 (**e**). Node heights in each tree are the median values obtained from Bayesian phylogenetic analysis with BEAST. The MRCA of each lineage is labeled with inferred dates. Scale bar, 0.08 substitutions per variable site. Two isolates have been removed from L5 and one isolate has been removed from L2 for the BEAST analyses.

© 2014 Nature America, Inc. All rights reserved.