AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

OBSERVATION

Recent Genetic Changes Affecting Enterohemorrhagic *Escherichia coli* Causing Recurrent Outbreaks

Joshua L. Cherry[a,b]

[a]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA
[b]Division of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, Bethesda, Maryland, USA

ABSTRACT Enterohemorrhagic *E. coli* (EHEC) is responsible for significant human illness, death, and economic loss. The main reservoir for EHEC is cattle, but plant-based foods are common vectors for human infection. Several outbreaks have been attributed to lettuce and leafy green vegetables grown in the Salinas and Santa Maria regions of California. Bacteria causing different outbreaks are mostly not close relatives, but one group of closely-related O157:H7 has caused several of them. This unusual pattern of recurrence may have some genetic basis. Here I use whole-genome sequences to reconstruct the genetic changes that occurred in the recent ancestry of this EHEC. In a short period of time corresponding to little genetic change, there were several changes to adhesion-related sequences, mainly adhesins. These changes may have greatly altered the adhesive properties of the bacteria. Possible consequences include increased persistence of cattle infections, more bacteria shed in cattle feces, and greater virulence in humans. Similar constellations of genetic change, which are detectable by current sequencing-based surveillance, may identify other bacteria that are particular threats to human health. In addition, the Santa Maria subclade carries a nonsense mutation affecting ArsR, a repressor of genes that confer resistance to arsenic and antimony. This suggests that the persistent source of Santa Maria contamination is located in an area with arsenic-contaminated groundwater, a problem in many parts of California. This inference may aid identification of the reservoir of EHEC, which would greatly aid mitigation efforts.

IMPORTANCE Food-borne bacterial infections cause substantial illness and death. Understanding how bacteria contaminate food and cause disease is important for combating the problem. Closely-related *E. coli*, likely originating in cattle, have repeatedly caused outbreaks spread by vegetables grown in California. Such recurrence is atypical, and might have a genetic basis. The genetic changes that occurred in the recent ancestry of these *E. coli* can be reconstructed from their DNA sequences. Several mutations affect genes involved in bacterial adhesion. These might affect persistence of infection in cattle, quantity of bacteria in their feces, and human disease. They also suggest a way of detecting dangerous bacteria from their genome sequences. Furthermore, a subgroup carries a mutation affecting the regulation of genes conferring arsenic resistance. This suggests that the reservoir for contamination utilizes groundwater contaminated with arsenic, a problem in parts of California. This observation may be an aid to locating the persistent reservoir of contamination.

KEYWORDS *Escherichia coli*, adhesins, adhesion molecules, arsenic resistance, food-borne pathogens

Enterohemorrhagic *E. coli* (EHEC) cause substantial human morbidity and mortality. They produce Shiga toxin and are most commonly of serotype O157:H7 (1–3). Their reservoir is mainly cattle, which, unlike humans, do not experience severe symptoms as a result of infection. Some strains can colonize the rectoanal junction (RAJ) of cattle (4), leading to a

Editor John M. Atack, Griffith University

Address correspondence to jcherry@ncbi.nlm.nih.gov.
The authors declare no conflict of interest.

Received 10 February 2022
Accepted 25 March 2022
Published 25 April 2022

persistent infection and in some cases a "supershedder" phenomenon (5, 6), in which orders of magnitude more bacteria are excreted in the feces.

Lettuce and leafy green vegetables grown in the Salinas and Santa Maria Valleys of California have been vectors for several outbreaks of EHEC in recent years. Though most have been caused by O157:H7 strains, in most cases strains causing different outbreaks have not been otherwise very closely related. Some of them, however, were recurrent outbreaks of very closely related O157:H7 (7, 8), differentiated by just a few single-nucleotide polymorphisms (SNPs) in the entire ~5Mb genome.

Although the *E. coli* causing the recurrent outbreaks from Salinas are distinguishable from those from Santa Maria, only a few SNPs separate them. This suggests that genetic traits of the bacteria contribute to the recurrence of contamination or to the human infection rate or symptom severity.

The NCBI Pathogens database contains clusters of closely-related isolates of foodborne pathogen species, including *E. coli*. It also provides a phylogenetic tree for each cluster, and information about single-nucleotide polymorphisms (SNPs) within the cluster. The history of nucleotide changes along the branches of the tree can be inferred by maximum parsimony reconstruction of ancestral states (9, 10). The effects of nucleotide changes on proteins can be determined because their location in the genome is known, and the genome is annotated with the coding sequences. Details of these procedures are given in Text S1 in the supplemental material.

Fig. 1 shows the phylogenetic tree for the cluster of *E. coli* containing the recurrent outbreak isolates (PDS000035073.159). The clade of isolates attributed to Salinas is shown in blue, and those attributed to Santa Maria are shown in red. Together with one other isolate, these form a larger clade. The isolates in the tree are all very closely related, differing by at most 69 SNPs.

The uncolapsed tree is shown in Fig. S1 in the supplemental material. Information about the isolates can be found at the NCBI Pathogen Detection website (PDT000639468.2).

Adhesion-related mutations. The line of descent from the root to the last common ancestor of the Salinas and Santa Maria clades is colored green in Fig. 1. Eighteen single-nucleotide changes that alter protein sequences are inferred to have occurred along this path: 2 nonsense mutations, and 16 that change one amino acid to another.

Of these 18 mutations, 4 affect adhesins (Fig. 1, Table 1). Two are in the same gene and occur on the same branch of the tree. The gene encodes an immunoglobulin-like domain-containing protein, which contains several domains of types associated with adhesins. The others are a nonsense mutation in *fdeC* and a nonsynonymous mutation in *bigA*.

In addition to these point mutations, a deletion of 1,218 bp (406 amino acids) occurred in the adhesin *yeeJ*, which mediates adhesion to abiotic surfaces and promotes biofilm formation (11, 12). The deletion is apparently the result of recombination between two copies of a 14 bp sequence in the gene. It eliminates 4 of the 17 immunoglobulin-like domains. The ancestral gene is similar in length to the longer variant previously described (12), but the deletion allele is distinct from the shorter variant and therefore might differ functionally.

Adhesins are found on the surface of bacteria and mediate adhesion to host cells, abiotic surfaces, and/or other bacterial cells, and can promote biofilm formation. Adhesins are involved in EHEC adherence to the cells of the RAJ (13). The adhesin mutations described above may alter their adhesive properties, enabling colonization of the RAJ or otherwise accounting for the recurrence of outbreaks. They might also contribute to human pathogenesis, though this effect would not be subject to selection because human infections do not usually spread far beyond the individual first infected.

The nonsense mutation in *fdeC* likely does not inactivate it. It occurs at position 1,177 of a 1,417 amino acid protein with many domains. An even shorter version of this protein binds to mammalian cells (14). Truncation may even be necessary for this binding (15). In any case, inactivation of adhesin genes can promote adhesion to cells of the RAJ (13).

A nonsense mutation also occurred in *arpA*, which encodes an ankyrin repeat protein (erroneously annotated in the MG1655 genome [16]). Deletion of *arpA* is associated with neonatal meningitis (17). The protein is distantly related to EspL2, an enterotoxin that promotes adhesion (18, 19), which suggests that this nonsense mutation affects adhesion.

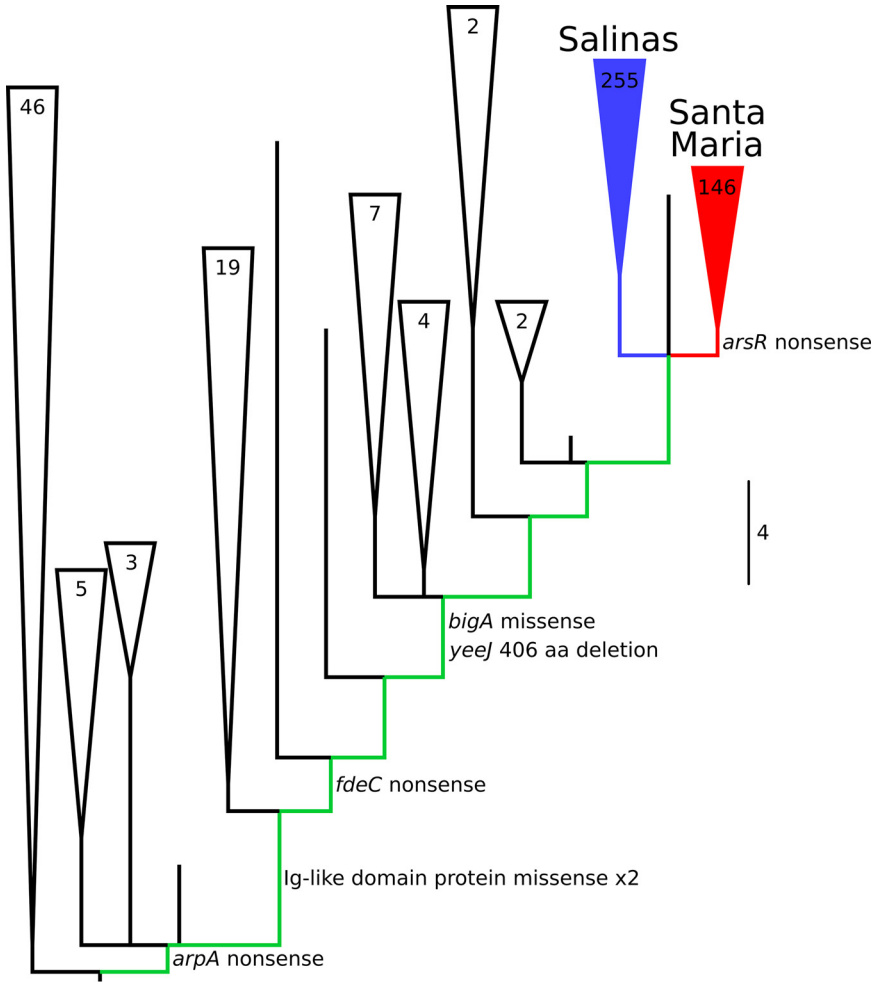

FIG 1 Phylogenetic tree of the cluster containing the Salinas and Santa Maria recurrent outbreak strains. Triangles represent collapsed clades, and the number within each indicates the number of isolates in the clade. The total number of isolates is 494. Mutations of interest are indicated to the right of the branches on which they occurred. The scale bar corresponds to four SNPs. The height of the tree from the root to the most distant tip is 36 SNPs. The tree obtained from the NCBI Pathogens data was rerooted based on outgroup sequences.

Constitutive expression of arsenic resistance genes. A nonsense mutation in *arsR* occurred along the branch leading to the last common ancestor of the Santa Maria clade. This gene encodes the repressor of *arsB* and *arsC*, the expression of which confers resistance to arsenic and antimony. This mutation truncates the 117 amino acid protein to 88 residues. Truncation of a similar ArsR (95% identical) to 89 residues causes increased expression of a reporter gene in the absence of inducer, corresponding to about 20% of the induced level (20). Truncation to 87 residues abolishes repressor dimerization, which is apparently necessary for repression (20). The nonsense mutation is therefore expected to result in at least significant partial induction, and perhaps full induction, of the arsenic resistance genes in the absence of inducer.

This mutation would likely be deleterious in the absence of arsenic or antimony, if only because of wasteful gene expression; this is the presumptive reason for the existence of the repressor. In the presence of high concentrations of arsenic or antimony, the mutation would not be deleterious, since the resistance genes would be fully expressed with or without it. It might, in fact, confer an advantage in an environment in which arsenic or antimony is present intermittently, as it would diminish or eliminate phenotypic lag for resistance.

Arsenic in ground water is a problem in many regions, including parts of California. The problem is greatest in portions of the Central Valley (21–23), to the east and north of Santa Maria. This area is the location of extensive cattle operations, and might contain

TABLE 1 Mutations discussed in the text[a]

Gene name[b]	Protein description	Nucleotide sequence identifier	Nucleotide position	Nucleotide change	Protein sequence identifier	Protein position	Protein change[b]	Protein length (ancestral)
-	Ig-like domain-containing protein	AAQXIR010000008.1	234093	C→T	EEQ5358295.1	212	A→V	1461
-	Ig-like domain-containing protein	AAQXIR010000008.1	234346	T→G	EEQ5358295.1	296	F→L	1461
fdeC	Intimin-like adhesin	AAQXIR010000008.1	9658	C→A	EEV2727796.1	1177	S→*	1417
bigA	Surface-exposed virulence protein	AAQXIR010000021.1	61345	C→A	EEQ5359752.1	150	P→Q	1011
yeeJ	Inverse autotransporter adhesin	AARLPB010000083.1	5978-7195	Deletion	EEV2728103.1	1036-1441	Deletion	2660
arpA	ShET2/EspL2 family type III secretion system effector toxin	AAQXIR010000007.1	279847	G→T	EEV2726747.1	133	E→*	728
arsR	Arsenic/antimony resistance repressor	AAQXIR010000005.1	121773	G→A	EEQ5357400.1	89	W→*	117

[a]The locations of the mutations on the nucleotide and protein sequences and their nature are described. Most of the sequence identifiers are for the reference sequence for the cluster, that of PDT000639468.2, a member of the Salinas clade. The exceptions are the protein sequences for *fdeC* and *arpA* (because protein sequences do not exist for the reference, which has premature stop codons) and the sequences for *yeeJ* (so that the full-length ancestral coding sequence and protein can be listed). These are taken from isolate PDT000430557.2, which is from the other side of the bifurcation at the root of the tree. The reference isolate has the derived allele for all genes but *arsR*. The protein lengths reported are for the ancestral protein sequences, which are longer than the derived sequences in the cases of nonsense mutation and deletion.
[b]"-" indicates that no gene name applies. An asterisk ("*") indicates a stop codon at the corresponding nucleotide positions.

the persistent source of *E. coli* responsible for recurrent Santa Maria outbreaks. The *E. coli* could make its way to Santa Maria in manure used for fertilizer or through movement of cattle. Some sources report high arsenic levels in a smaller fraction of wells in the Santa Maria Valley (23). A search for a local reservoir could focus on cattle operations using water from wells with high arsenic levels.

Conclusion. Changes to adhesion-related proteins likely play a role in the recurring EHEC outbreaks vectored by vegetables grown in Salinas and Santa Maria valleys. The combination of several such mutations may have radically altered the adhesive properties of the bacteria and the nature of host colonization. In addition to contributing to persistence and shedding in cattle, they might increase pathogenicity in humans, leading to a greater probability of outbreak detection and the reporting of a higher fraction of cases. Altered adhesive properties may play a more general role in foodborne outbreaks. The occurrence of multiple changes to adhesion-related sequences in a short time may indicate that the affected bacteria pose a particular threat to human health.

The *arsR* nonsense mutation that affects the Santa Maria clade may be a clue to the location of the source of repeated contamination. The use of sequences for tracing sources of infection is usually based on the inferred relationships between isolates. The case of *arsR* exemplifies a different kind of inference, in which the expected phenotypic effects of mutations have implications for the environment in which a source of infections is growing.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.1 MB.

ACKNOWLEDGMENTS

I thank David Lipman for invaluable discussions. This work was supported by the intramural research program of the National Library of Medicine, National Institutes of Health. The opinions expressed in this article are those of the author and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

REFERENCES

1. Beutin L. 2006. Emerging enterohaemorrhagic *Escherichia coli*, causes and effects of the rise of a human pathogen. J Vet Med B Infect Dis Vet Public Health 53:299–305. https://doi.org/10.1111/j.1439-0450.2006.00968.x.
2. Nguyen Y, Sperandio V. 2012. Enterohemorrhagic *E. coli* (EHEC) pathogenesis. Front Cell Infect Microbiol 2:90. https://doi.org/10.3389/fcimb.2012.00090.
3. Croxen MA, Finlay BB. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. Nat Rev Microbiol 8:26–38. https://doi.org/10.1038/nrmicro2265.
4. Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ, Pearce MC, McKendrick IJ, Smith DGE, Gally DL. 2003. Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia*

coli O157:H7 in the bovine host. Infect Immun 71:1505–1512. https://doi.org/10.1128/IAI.71.3.1505-1512.2003.
5. Arthur TM, Brichta-Harhay DM, Bosilevac JM, Kalchayanand N, Shackelford SD, Wheeler TL, Koohmaraie M. 2010. Super shedding of *Escherichia coli* O157:H7 by cattle and the impact on beef carcass contamination. Meat Sci 86:32–37. https://doi.org/10.1016/j.meatsci.2010.04.019.
6. Cobbold RN, Hancock DD, Rice DH, Berg J, Stilborn R, Hovde CJ, Besser TE. 2007. Rectoanal junction colonization of feedlot cattle by *Escherichia coli* O157:H7 and its association with supershedders and excretion dynamics. Appl Environ Microbiol 73:1563–1568. https://doi.org/10.1128/AEM.01742-06.
7. U.S. Food & Drug Administration. 2020. Outbreak investigation of *E. coli*: romaine from Salinas, California (November 2019). https://www.fda.gov/food/outbreaks-foodborne-illness/outbreak-investigation-e-coli-romaine-salinas-california-november-2019. Retrieved 10 January 2022.
8. U.S. Food & Drug Administration. 2021. Outbreak investigation of *E. coli*—leafy greens (December 2020). https://www.fda.gov/food/outbreaks-foodborne-illness/outbreak-investigation-e-coli-leafy-greens-december-2020. Retrieved 10 January 2022.
9. Cherry JL. 2018. Methylation-induced hypermutation in natural populations of bacteria. J Bacteriol 200:e00371-18. https://doi.org/10.1128/JB.00371-18.
10. Cherry JL. 2020. Selection-driven gene inactivation in Salmonella. Genome Biol Evol 12:18–34. https://doi.org/10.1093/gbe/evaa010.
11. Roux A, Beloin C, Ghigo J-M. 2005. Combined inactivation and expression strategy to study gene function under physiological conditions: application to identification of new *Escherichia coli* adhesins. J Bacteriol 187:1001–1013. https://doi.org/10.1128/JB.187.3.1001-1013.2005.
12. Martinez-Gil M, Goh KGK, Rackaityte E, Sakamoto C, Audrain B, Moriel DG, Totsika M, Ghigo J-M, Schembri MA, Beloin C. 2017. YeeJ is an inverse autotransporter from *Escherichia coli* that binds to peptidoglycan and promotes biofilm formation. Sci Rep 7:11326. https://doi.org/10.1038/s41598-017-10902-0.
13. Moreau MR, Kudva IT, Katani R, Cote R, Li L, Arthur TM, Kapur V. 2021. Nonfimbrial adhesin mutants reveal divergent *Escherichia coli* O157:H7 adherence mechanisms on human and cattle epithelial cells. Int J Microbiol 2021:8868151. https://doi.org/10.1155/2021/8868151.
14. Nesta B, Spraggon G, Alteri C, Moriel DG, Rosini R, Veggi D, Smith S, Bertoldi I, Pastorello I, Ferlenghi I, Fontana MR, Frankel G, Mobley HLT, Rappuoli R, Pizza M, Serino L, Soriani M. 2012. FdeC, a novel broadly conserved *Escherichia coli* adhesin eliciting protection against urinary tract infections. mBio 3:e00010-12. https://doi.org/10.1128/mBio.00010-12.
15. Easton DM, Allsopp LP, Phan M-D, Moriel DG, Goh GK, Beatson SA, Mahony TJ, Cobbold RN, Schembri MA. 2014. The intimin-like protein FdeC is regulated by H-NS and temperature in enterohemorrhagic *Escherichia coli*. Appl Environ Microbiol 80:7337–7347. https://doi.org/10.1128/AEM.02114-14.
16. Wolfe AJ. 2005. The acetate switch. Microbiol Mol Biol Rev 69:12–50. https://doi.org/10.1128/MMBR.69.1.12-50.2005.
17. Clermont O, Bonacorsi S, Bingen E. 2004. Characterization of an anonymous molecular marker strongly linked to *Escherichia coli* strains causing neonatal meningitis. J Clin Microbiol 42:1770–1772. https://doi.org/10.1128/JCM.42.4.1770-1772.2004.
18. Miyahara A, Nakanishi N, Ooka T, Hayashi T, Sugimoto N, Tobe T. 2009. Enterohemorrhagic *Escherichia coli* effector EspL2 induces actin microfilament aggregation through annexin 2 activation. Cell Microbiol 11:337–350. https://doi.org/10.1111/j.1462-5822.2008.01256.x.
19. Tobe T. 2010. Cytoskeleton-modulating effectors of enteropathogenic and enterohemorrhagic *Escherichia coli*: role of EspL2 in adherence and an alternative pathway for modulating cytoskeleton through Annexin A2 function. FEBS J 277:2403–2408. https://doi.org/10.1111/j.1742-4658.2010.07654.x.
20. Xu C, Rosen BP. 1997. Dimerization is essential for DNA binding and repression by the ArsR metalloregulatory protein of Escherichia coli. J Biol Chem 272:15734–15738. https://doi.org/10.1074/jbc.272.25.15734.
21. The Environmental Integrity Project. 2016. Arsenic in California drinking water. http://environmentalintegrity.org/wp-content/uploads/CA-Arsenic-Report.pdf. Retrieved 12 August 2021.
22. Haugen EA, Jurgens BC, Arroyo-Lopez JA, Bennett GL. 2021. Groundwater development leads to decreasing arsenic concentrations in the San Joaquin Valley, California. Sci Total Environ 771:145223. https://doi.org/10.1016/j.scitotenv.2021.145223.
23. Smith R, Knight R, Fendorf S. 2018. Overpumping leads to California groundwater arsenic threat. Nat Commun 9:2089. https://doi.org/10.1038/s41467-018-04475-3.