

### ABSTRACT

Enterohemorrhagic *Escherichia coli* (EHEC) are a leading cause of *E. coli*-related foodborne illness in the United States. Most characterized members of this pathotype belong to the O157:H7 serotype and are part of the EHEC 1 clonal group. Another group of strains, EHEC 2, is capable of causing a clinically indistinguishable disease. Extensive genomic plasticity has been observed among EHEC 1 isolates, but little is known about the degree of chromosomal variability among EHEC 2 isolates. Optical mapping was performed on 26 EHEC 2 strains to uncover insertions and deletions (indels) larger than 2 kb in length as well as chromosomal rearrangements. Serotypes investigated included O26:H11 (n=10), O111:H8 (n=10), and O118:H16 (n=6). Optical maps were compared to *in silico* maps of the EHEC 2 genome reference strains 11368 (O26:H11) and 11128 (O111:H8). The O26:H11 and O118:H16 strains possessed, on average, larger chromosomes than the O111:H8 strains (5.7 vs. 5.4 Mb). Most of the observed map variation was within regions known to contain prophages and other integrative elements. Comparative analyses identified 265 markers for strain identification and subtyping purposes. Of these, 239 were indels, 22 were inversions, and 4 were suggestive of double crossover events. As a group, the O26:H11 strains were the most variable with an average map similarity of 93.8%. Lower levels of within serotype variation were observed for both the O111:H8 and O118:H16 strains with these serotypes having average map similarities of 96.2% and 97.0%, respectively. Cluster analysis of the map polymorphisms resolved each of the serotypes with the O26:H11 and O118:H16 strains more similar to each other than either was to O111:H8. Indel and rearrangement events, dominated by prophages, have played a large role in shaping the chromosomal architecture of this group of closely related pathogens during their evolution and diversification.

### INTRODUCTION

Enterohemorrhagic *Escherichia coli* (EHEC) are a leading cause of *E. coli*-related foodborne illness in the United States. This group of pathogenic *E. coli* causes a variety of human and animal illnesses ranging from diarrhea to hemorrhagic colitis, and the multifactorial hemolytic uremic syndrome (3). The most well-characterized members of this pathotype belong to the O157:H7 serotype and are part of the EHEC 1 clonal group (7). Another group of strains, EHEC 2, is capable of causing a clinically indistinguishable disease. Serotypes within this group include O26:H11, O111:H8, and O118:H16, as well as non-motile variants (1, 7). Substantial genomic plasticity has been observed among EHEC 1 isolates (4), but little is known about the degree of chromosome structural variation among EHEC 2 isolates. Optical mapping was performed on 26 EHEC 2 strains and comparative analyses were conducted using the complete genomes of two additional EHEC 2 isolates (6) to uncover insertions and deletions (indels) larger than 2 kb in length as well as chromosomal rearrangements.

### MATERIALS & METHODS

**Optical mapping:** Twenty-six bacterial isolates were optically mapped by digesting surface-bound chromosomes with BamHI and the contiguous restriction fragments were measured with a CCD camera (OpGen, Gaithersburg, MD). Multiple overlapping contigs were assembled to create a restriction map of the entire circular chromosome for each strain. Backbone sequences were generated for the EHEC 2 genomic reference strains 11368 and 11128 (6) by removing sequences associated with the annotated prophages, integrative elements, and insertion sequences. Optical maps were compared to the BamHI restriction maps of the backbone sequences, and indel and rearrangement events were characterized.

**Phylogenetic analyses:** Genome similarity was calculated by dividing the sum of the aligned restriction fragment lengths by the sum of the total map lengths for all pairwise comparisons of strains. The resulting distance matrix was then used for phylogenetic analysis with the neighbor joining algorithm as implemented by MEGA 3 (5). The presence or absence of indels and rearrangements was converted to binary data and then imported into SplitsTree 4 (2) for neighbor-net analysis using the uncorrected *p* distance.

TABLE 1. Characteristics of strains investigated

Group	Strain	pEHEC										Isolation data		Map length (bp)	Inv.*	DCO**			
		eae	stx1	stx2	hly	hlyA	hlyB	hlyC	hlyD	hlyE	hlyF	Host	Locate, Year						
O26:H11	11368	+	+	+	+	+	+	+	+	+	+	+	+	+	human	Japan, 2001	5,697,240	1	1
	EC0400	+	+	+	+	+	+	+	+	+	+	+	+	+	no data		5,665,113	2	0
	EC1649	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (ID), 1987	5,897,861	2	1
	EC1651	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Germany, 1998	5,673,405	5	0
	EC1664	+	+	+	+	+	+	+	+	+	+	+	+	+	human	Germany, 2000	5,818,177	1	1
	EC1741	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (SD), 1974	5,497,907	0	0
	EC1743	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	USA (SD), 1989	5,644,001	1	1
	EC1750	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (NE), 1998	5,545,112	3	1
	EC1753	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (DC), 1999	5,556,421	1	1
	EC1758	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (MI), 2003	5,607,990	3	0
EC1783	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Germany, 1998	5,761,964	1	0	
O111:H8	11128	+	+	+	+	+	+	+	+	+	+	+	+	+	human	Japan, 2001	5,371,077	1	0
	EC1370	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (ID), 1986	5,234,975	2	0
	EC1618	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (MD), 1977	5,324,304	1	0
	EC1631	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (WA), 1991	5,368,239	0	0
	EC1655	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (TX), 1999	5,327,624	1	0
	EC1668	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (WA), 2001	5,423,501	2	0
	EC1746	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Scotland, 1993	5,327,951	4	0
	EC1747	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	USA (CA), 1993	5,538,548	2	0
	EC1751	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (NE), 1998	5,487,580	1	0
	EC1752	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (NE), 1998	5,385,119	2	0
EC1755	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (WA), 1999	5,284,812	2	0	
O118:H16	EC1639	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Germany, 1989	5,686,257	1	0
	EC1669	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (WA), 2001	5,643,714	1	0
	EC1748	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Germany, 1994	5,640,918	0	0
	EC1749	+	+	+	+	+	+	+	+	+	+	+	+	+	bovine	Germany, 1994	5,831,872	0	0
	EC1754	+	+	+	+	+	+	+	+	+	+	+	+	+	human	Germany, 1996	5,692,154	2	0
	EC1756	+	+	+	+	+	+	+	+	+	+	+	+	+	human	USA (WA), 2000	5,673,968	1	0

\* Inv., inversion; DCO, double crossover.

FIGURE 1. Alignment of linearized *in silico* BamHI restriction maps of EHEC 2 reference genome strains 11368 and 11128. Vertical black lines within each map represent BamHI cut sites. Aligned fragments are shown in green, while unaligned ones are white. Contiguous aligned fragments are indicated by the lines connecting the maps. Chromosomal inversions are indicated by the dark green fragments. Annotated prophages (P) and integrative elements (IE) are shown by the yellow and blue boxes, respectively, and their directionality is indicated by the arrows.

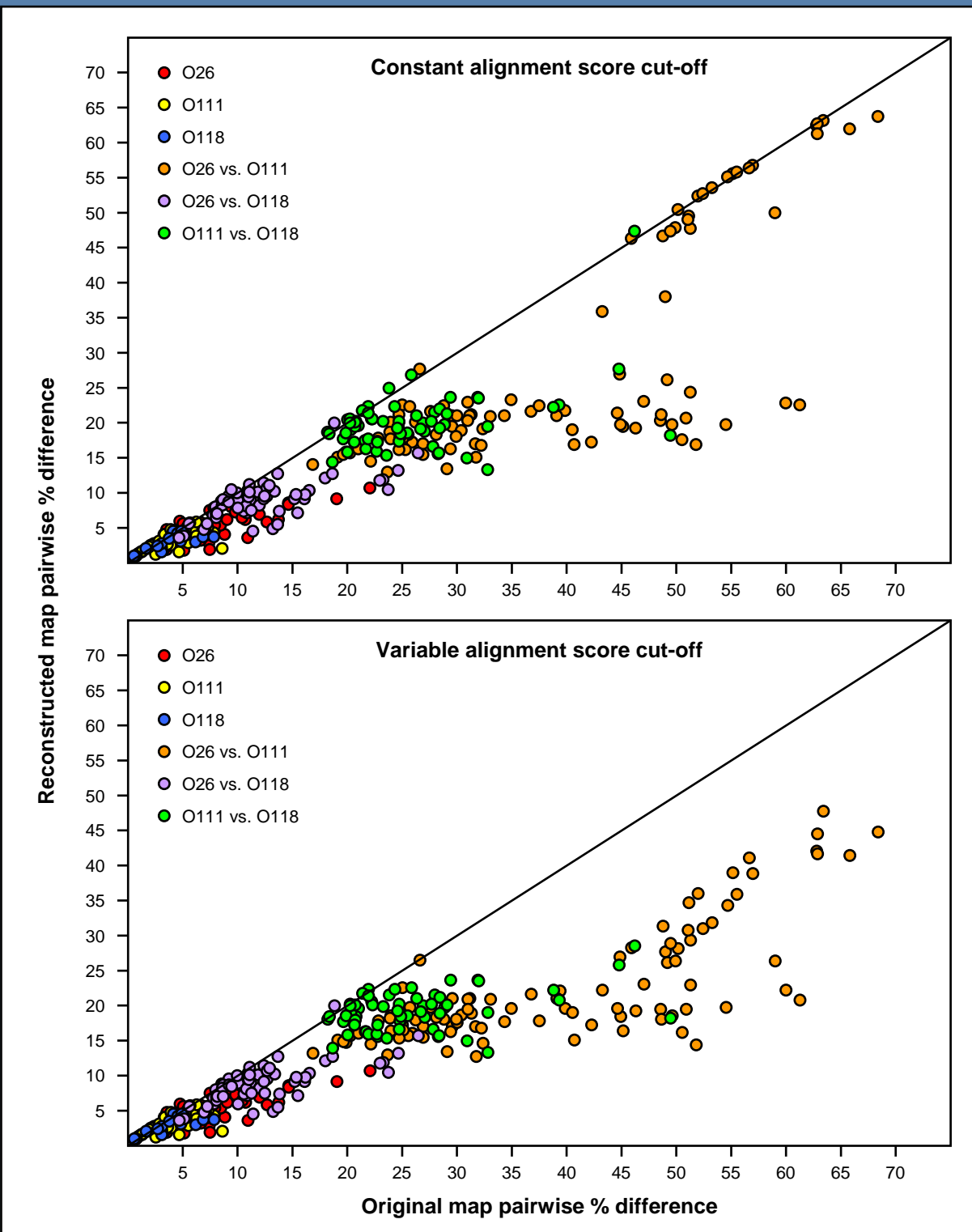
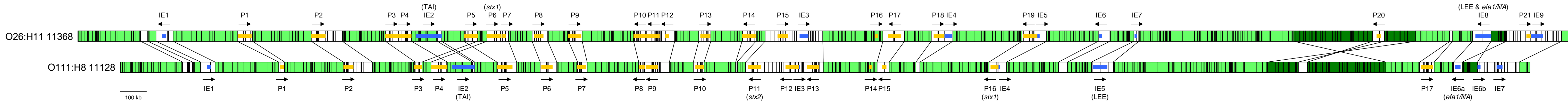


FIGURE 2. Pairwise map similarity improvement. *In silico* correction of identified chromosomal rearrangements significantly improved map similarity values (top panel). Further improvement was observed when a variable alignment score cut-off was used (bottom panel).

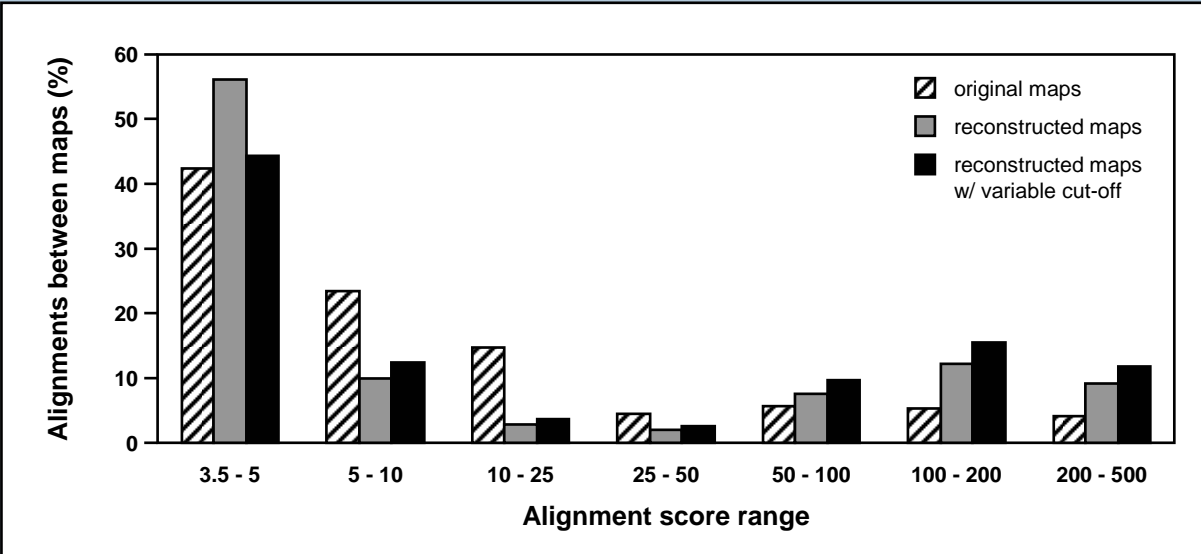


FIGURE 3. Alignment score distribution. *In silico* correction of identified chromosomal rearrangements and the use of a variable alignment score cut-off alter the observed alignment score distribution.

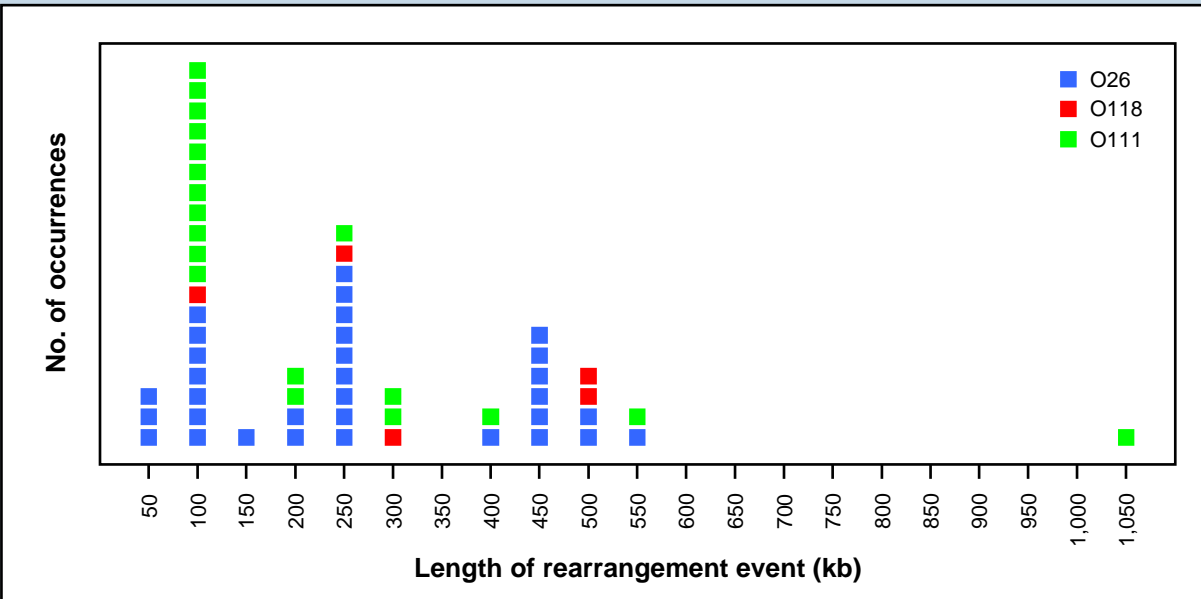


FIGURE 4. Chromosomal rearrangements. Most rearrangement events were between 50 and 550 kb. One O111:H8 strain (EC1752), however, possesses an inversion greater than 1 Mb in length.

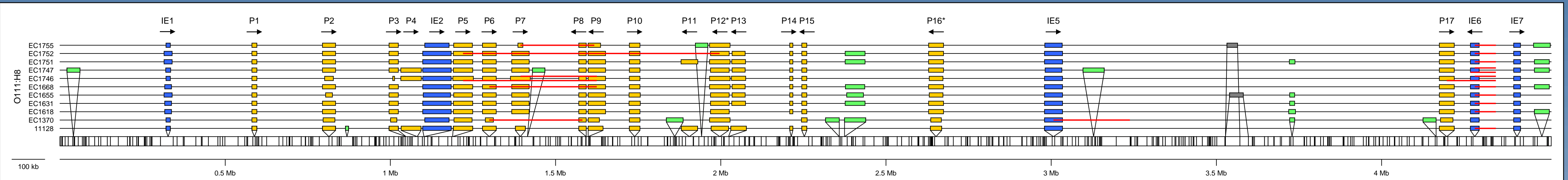


FIGURE 5. Locations of O111:H8 indels and rearrangements. Optical maps were compared to the linearized BamHI restriction map of the 11128 reference backbone sequence, which is shown as a white box with vertical black lines. Insertions corresponding to the locations of the known prophages (P) and integrative elements (IE) are colored yellow and blue, respectively. Other insertions are in green and deletions are in gray. Inversions are displayed as red lines. \*IE3 and IE4 cannot be distinguished from their associated prophages (P12 and P16, respectively) in the optical maps.

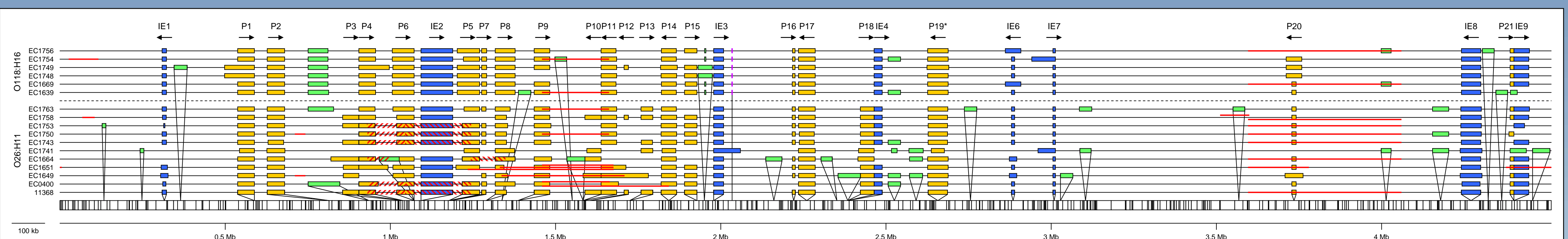


FIGURE 6. Locations of O26:H11 & O118:H16 indels and rearrangements. Optical maps were compared to the linearized BamHI restriction map of the 11368 reference backbone sequence, which is shown as a white box with vertical black lines. Insertions corresponding to the locations of the known prophages (P) and integrative elements (IE) are colored yellow and blue, respectively. Other insertions are in green and deletions are in gray. Inversions are displayed as red lines, while the regions exchanged in probable double crossover events are shown by the red hatched boxes. \*IE5 cannot be distinguished from its associated prophage, P19, in the optical maps.

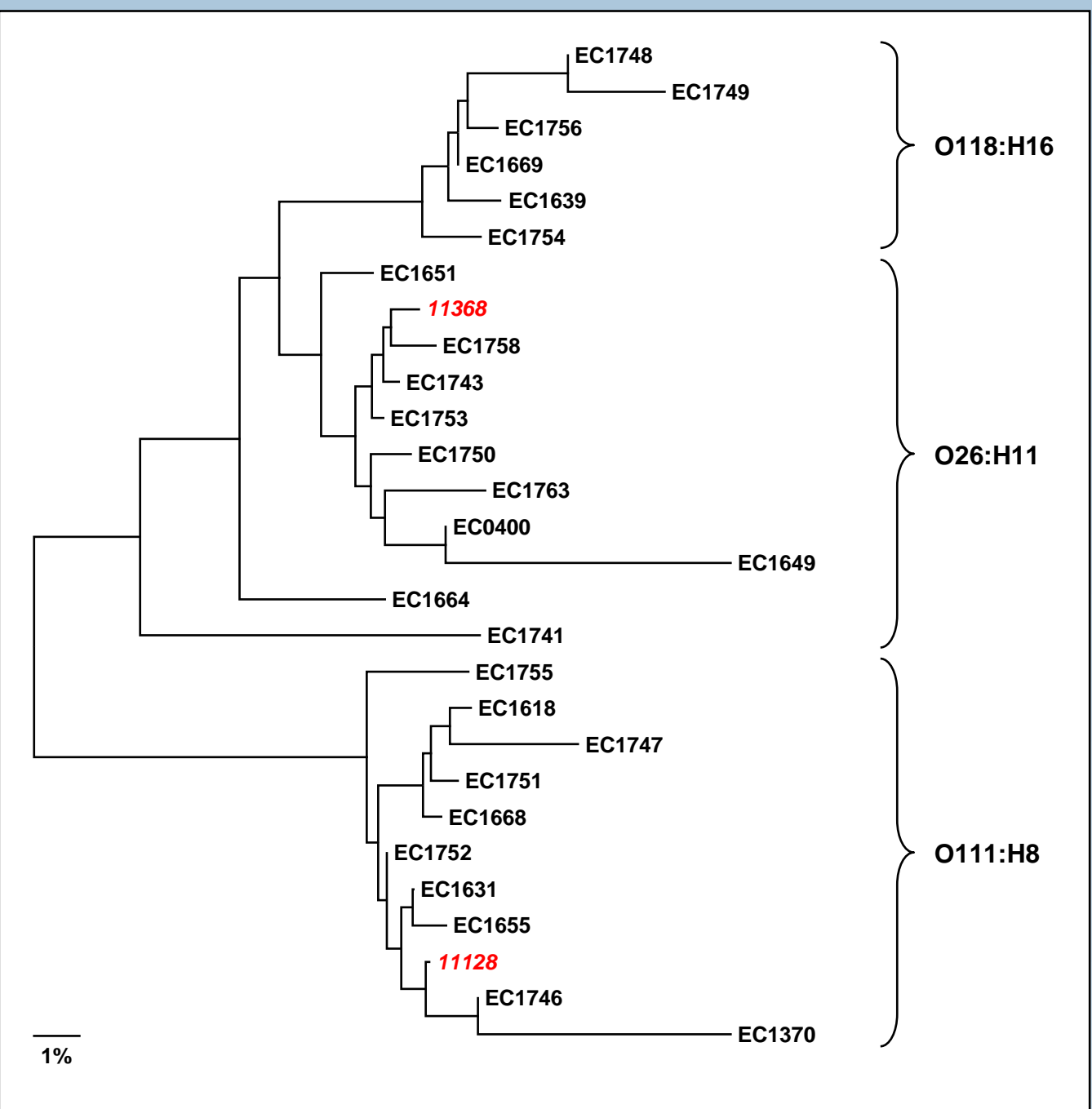


FIGURE 7. Genome similarity phylogenetic analysis. Chromosomal rearrangements were corrected *in silico* and a variable alignment score cut-off was used in generating the distance matrix. This matrix was then imported into MEGA for cluster analysis using the neighbor joining algorithm. Genome reference strains are shown in red italics.

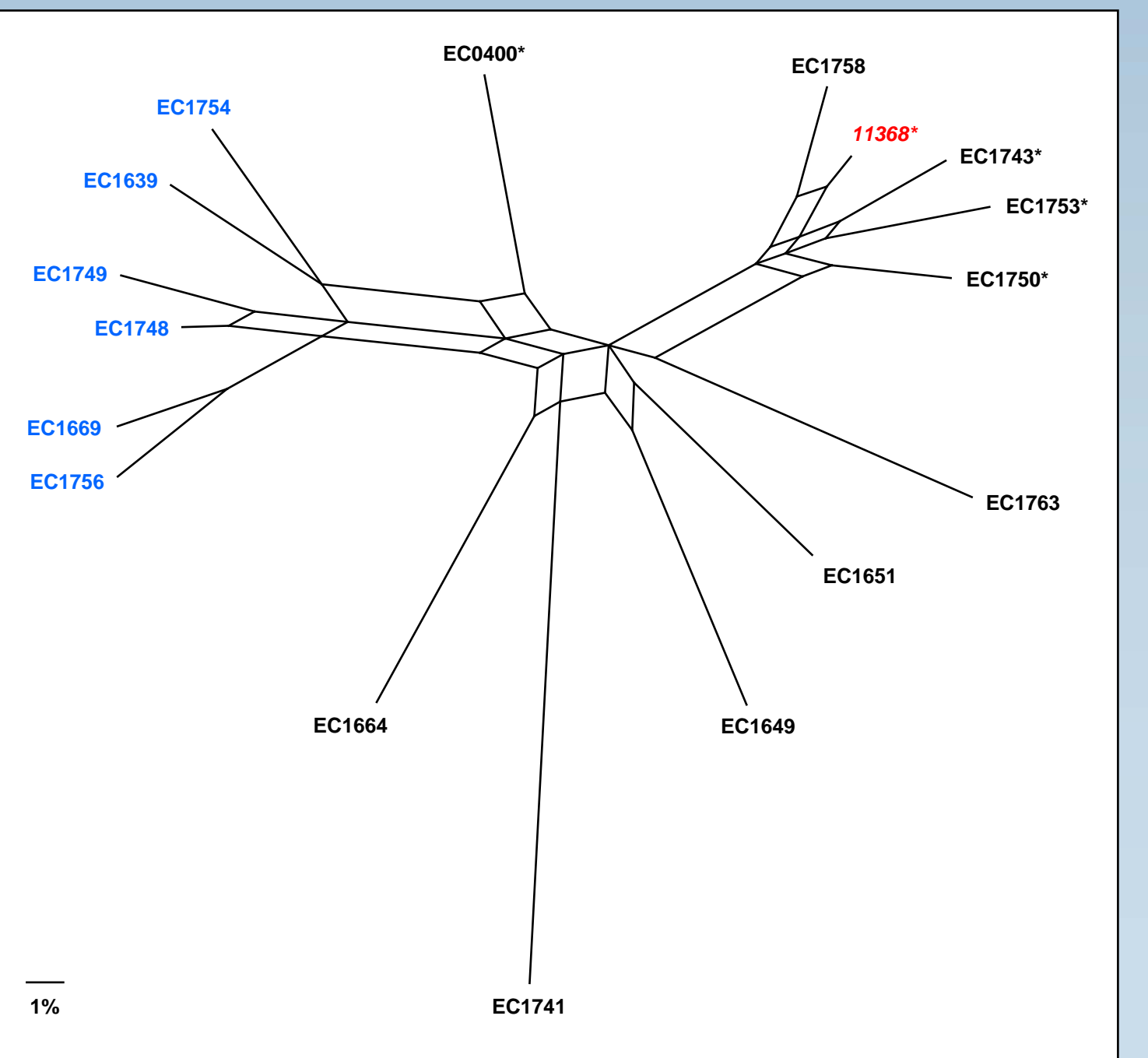


FIGURE 8. O26:H11 & O118:H16 phylogenetic network analysis. Indel and rearrangement map variants were converted into binary data and imported into SplitsTree for neighbor-net analysis using the *p* distance. The genome reference strain is shown in red italics, while the O118:H16 strains are in blue. O26:H11 strains possessing the P4/5/6 double crossover event are indicated by an asterisk.

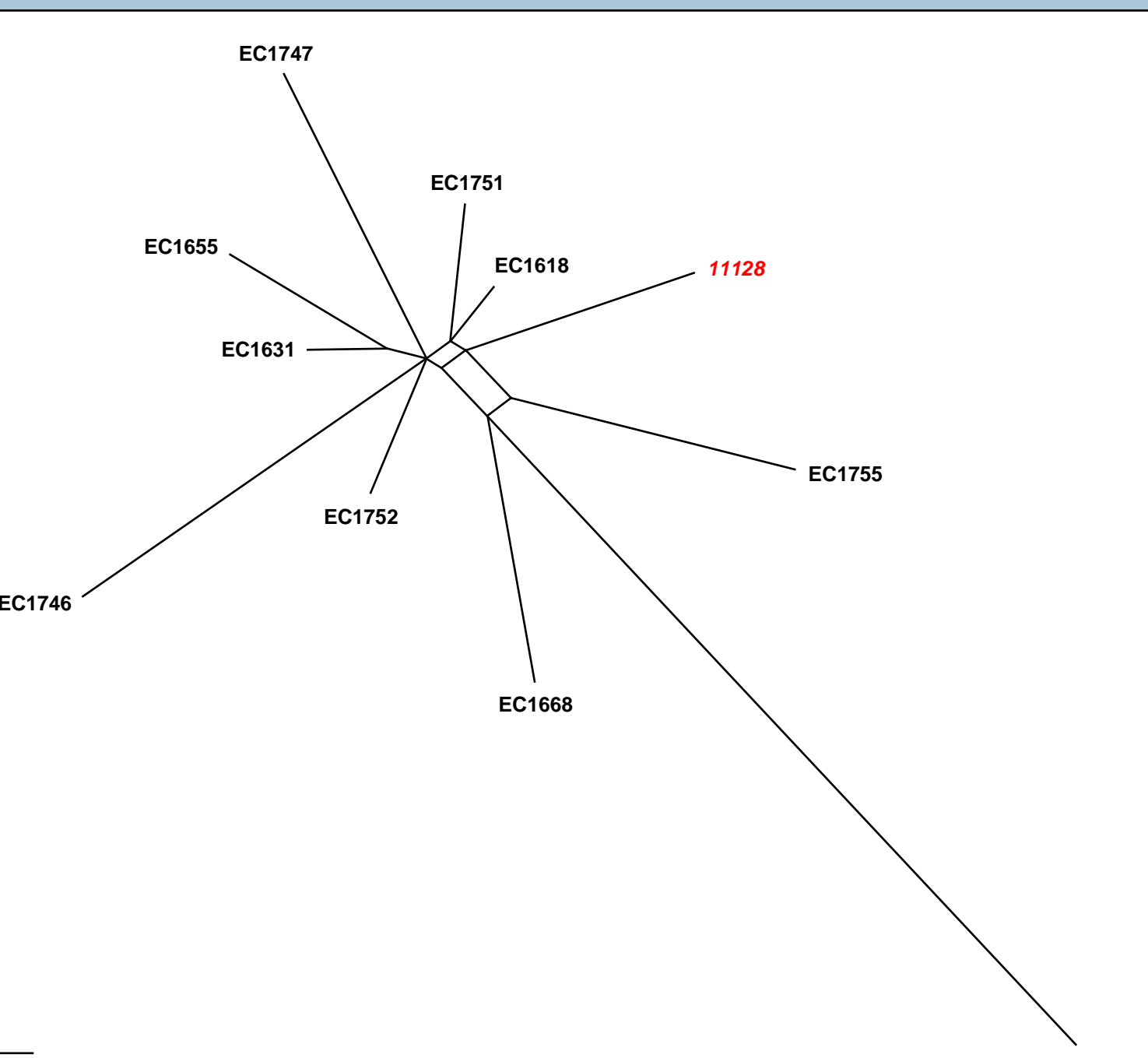


FIGURE 9. O111:H8 phylogenetic network analysis. Indel and rearrangement map variants were converted into binary data and imported into SplitsTree for neighbor-net analysis using the *p* distance. The genome reference strain is shown in red italics.

### RESULTS

#### Map alignment scores & cut-off values

Pairwise map similarities were improved through the *in silico* correction of chromosomal rearrangements and the use of variable alignment score cut-off values.

For within serogroup comparisons, an alignment score cut-off of 10 was determined to give the best results with no multiple alignments among the homologous prophages.

An alignment score cut-off of 5 was selected for the O26-O118 between group comparisons, while a cut-off of 3.5 was used for the O26-O111 and O118-O111 comparisons.

#### Group map similarity averages

Within group map similarity averages were 93.8, 96.2, and 97.0% for the O26, O111, and O118 strains, respectively. Exclusion of outlier strains (EC1370, EC1649, EC1664, and EC1741) raised the O26 and the O111 groups' average map similarities to 96.6 and 97.0%, respectively.

Between group map similarity averages were 77.9, 81.1, and 91.4% for the O26-O111, O111-O118, and O26-O118 group comparisons, respectively.

#### Chromosomal markers

40, 37, and 33 indel sites were identified among the O26, O111, and O118 strains, respectively. Map variation at these sites resulted in 15, 10, and 3 rearrangement markers for the O26, O111, and O118 strains, respectively.

19, 12, and 6 rearrangement sites were identified among the O26, O111, and O118 strains, respectively. Map variation at these sites resulted in 15, 10, and 3 rearrangement markers for the O26, O111, and O118 strains, respectively.

In total, 239 indel markers and 28 rearrangement markers were identified among the 28 strains investigated.

### CONCLUSIONS

Isolates belonging to the same serotype within a clonal group show, on average, 3-4% variation at the optical map level.

- Similar levels of map variation were observed within each EHEC 2 serotype. This is comparable to the amount of map variation observed within O157:H7 and within O55:H7 (~3% for EHEC 2 serotypes vs. ~4% for EHEC 1 serotypes).

In comparison to the EHEC 1 clonal group, higher levels of chromosomal plasticity were observed among the EHEC 2 isolates. This suggests that the EHEC 2 clonal group is more variable than the EHEC 1 group.

- Based on % map similarities, the EHEC 2 O26:H11 and O118:H16 serotypes are more closely related to each other than the EHEC 1 O157:H7 and O55:H7 serotypes are to each other (91.4% vs. 87.5%).

- However, the O111:H8 isolates examined are, on average, ~20% different from either the O26:H11 or O118:H16 isolates. This level of variation is greater than that observed between O157:H7 and O55:H7 (12.5%).

### FUNDING

This project has been funded in part by the National Biodefense Analysis Center of the Department of Homeland Security under Interagency Agreement Grant # 224-04-2806 and by the National Institute of Allergy and Infectious Diseases (NIAID), NIH, DHHS, under NIAID contract N01-AI-30071.

### REFERENCES

1. Abu-Ali GS, Lacher DW, Wick LM, Qi W, and Whitam TS. 2009. Genomic diversity of pathogenic *Escherichia coli* of the EHEC 2 clonal complex. *BMC Genomics* 10:206.
2. Hahn SH and Singer D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254-67.
3. Kaper JB, Nisalakorn S, and Monto AS. 2004. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* 2:22-34.
4. Kolesnik M, Marmann MK, LeClere JE, and Calzadilla TA. 2009. Optical mapping and 454 sequencing of *Escherichia coli* O157:H7 isolates linked to the US 2000 spinach-associated outbreak. *Microbiology* 154:261-8.
5. Kumar S, Tamura K, and Nei M. 2004. MEGA: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5:150-63.
6. Ogura Y, Ochiai T, Iguchi A, Tsuboi H, Aoshima M, Oshima K, Kozuma T, Abo H, Nakayama K, Kurakawa K, Tobe T, Hattori M, and Hayashi T. 2006. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc Natl Acad Sci U S A* 103:11000-5.
7. Reid SD, Heffern CJ, Burnough AC, Selander RK, and Whitam TS. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* 406:67-71.