

1 **Phylogeographic analysis reveals multiple international transmission events have driven the**
2 **global emergence of *Escherichia coli* O157:H7**

3
4 Short Title: The worldwide spread of *E. coli* O157:H7
5

6
7 **AUTHORS**

8
9 Eelco Franz¹, Ovidiu Rotariu², Bruno S. Lopes³, Marion MacRae³, James L. Bono⁴, Chad Laing⁵, Victor
10 Gannon⁵, Robert Söderlund⁶, Angela H.A.M. van Hoek¹, Ingrid Friesema¹, Nigel P. French⁷, Tessy
11 George⁷, Patrick J. Biggs⁷, Patricia Jaros⁷, Marta Rivas⁸, Isabel Chinen⁸, Josefina Campos⁸, Cecilia
12 Jernberg⁹, Kari Gobius¹⁰, Glen E. Mellor¹⁰, P. Scott Chandry¹⁰, Francisco Perez-Reche¹¹, Ken J. Forbes³
13 and Norval J.C. Strachan^{2*}
14

15 ¹National Institute for Public Health and the Environment (RIVM), Centre for Infectious Disease
16 Control (CIb), P.O. Box 1, 3720 BA Bilthoven, the Netherlands.

17 ²School of Biological Sciences, The University of Aberdeen, Cruickshank Building. St Machar Drive,
18 Aberdeen, Scotland, United Kingdom, AB24 3UU.

19 ³School of Medicine, Medical Sciences & Nutrition, The University of Aberdeen, Foresterhill,
20 Aberdeen, Scotland, United Kingdom, AB25 2ZD.

21 ⁴United States Department of Agriculture, Agricultural Research Service, US Meat Animal Research
22 Center, Clay Center, Nebraska.

23 ⁵National Microbiology Laboratory, Public Health Agency of Canada, 225089 Township Road 9-1
24 (Box 640), Lethbridge, Alberta, Canada, T1J 3Z4.

25 ⁶National Veterinary Institute (SVA), Uppsala, Sweden.

26 ⁷^mEpiLab, Infectious Disease Research Centre, School of Veterinary Science, Massey University,
27 Palmerston North, New Zealand.

28 ⁸INEI-ANLIS "Dr. Carlos G. Malbrán", Av. Vélez Sarsfield 563, (1281) Ciudad Autónoma de Buenos
29 Aires, Argentina.

30 ⁹Department of Microbiology, The Public Health Agency of Sweden, Stockholm, Sweden.

31 ¹⁰CSIRO Agriculture and Food, Werribee, Australia.

32 ¹¹Institute of Complex Systems and Mathematical Biology, SUPA, School of Natural and Computing
33 Sciences, University of Aberdeen, Aberdeen, Scotland, United Kingdom.

34

35 **Keywords:**

36 Infectious diseases, STEC, whole genome sequencing, phylogeography, E. coli O157

37

38 ***Corresponding author:**

39 Name: Prof. Norval Strachan

40 Address: School of Biological Sciences, The University of Aberdeen, Cruickshank

41 Building, St Machar Drive, Aberdeen, United Kingdom, AB24 3UU

42 Tel no: +44 1224 272699

43 Fax no: +44 1224 272703

44 Email address: n.strachan@abdn.ac.uk

45 **Alternate Corresponding Author:**

46 Name: Dr. Eelco Franz

47 Address: Head of Department Epidemiology and Surveillance of Enteric and Zoonotic

48 Infections, Centre for Infectious Disease Control (CIb), National Institute for

49 Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven,

50 the Netherlands

51 Tel.: +31 30 2747063

52 Fax: +31 6 15502479

53 Email address: eelco.franz@rivm.nl

54 **Word Count:** Abstract: 211 words, Text body: 2996 words

55

56 **Summary**

57

58 Phylogeographic analyses identified 34 major international transmission events, starting in
59 Europe around 1890, that resulted in the current distribution of *E. coli* O157:H7. These were
60 likely facilitated by global cattle movements and will inform policy to reduce this pathogens
61 spread.

62

63 **Abstract**

64

65 **Background.** Shiga toxin-producing *Escherchia coli* O157:H7 is a zoonotic pathogen which
66 causes numerous food and waterborne disease outbreaks. It is globally distributed but its origin
67 and temporal sequence of geographical spread is unknown.

68

69 **Methods.** We analysed Whole Genome Sequencing data of 757 isolates from 4 continents and
70 performed a pan genome analysis to identify the core genome and from this extracted single
71 nucleotide polymorphisms. Timed phylogeographic analysis was performed on a subset of the
72 isolates to investigate it's worldwide spread.

73

74 **Results.** The common ancestor of this set of isolates occurred around 1890 (1845–1925) and
75 originated from the Netherlands. Phylogeographic analysis identified 34 major transmission
76 events. The earliest were predominantly intercontinental from Europe to Australia around 1937
77 (1909-1958), to USA in 1941 (1921-1962), to Canada in 1960 (1943-1979), and from Australia
78 to New Zealand in 1966 (1943-1982). This pre-dates the first reported human case of *E. coli*
79 O157:H7 in 1975 from the USA.

80

81 **Conclusions.** Inter- and intra- continental transmission events have resulted in the current
82 international distribution of *E. coli* O157:H7 and it is likely that these events were facilitated
83 by animal movements (e.g. Holstein Friesian cattle). These findings will inform policy on
84 action that is crucial to reduce further spread of *E. coli* O157:H7 and other (emerging) STEC
85 strains globally.

86

87

88 **Introduction**

89

90 Emerging infectious diseases (EIDs) are a significant and growing threat to global health, economy and
91 security[1]. Most EIDs are driven by socio-economic, environmental and ecological factors[2].
92 Examples include the long-term spread and maintenance of cholera[3] and the (mis)use of
93 antimicrobials resulting in the emergence of multi-drug resistant bacteria[4]. Agricultural
94 intensification and/or environmental change have been associated with an increased risk of disease
95 emergence, driven by the impact of an expanding human population and changing human interaction
96 with the environment[5]. Animal trade has been identified as an underlying cause of the emergence and
97 spread of infectious disease as exemplified by avian influenza[6] and swine flu[7]. Understanding
98 emergence of disease is crucial in preventing such events in the future.

99

100 Shiga toxin-producing *Escherichia coli* (STEC) are globally dispersed zoonotic pathogens associated
101 with a broad spectrum of sequelae in humans, including diarrhoea, haemorrhagic colitis and
102 (occasionally fatal) haemolytic uremic syndrome (HUS)[8]. Cattle and sheep are generally considered
103 as the main reservoirs[9]. *E. coli* O157:H7 is the most commonly reported STEC serotype and was first
104 recognised as a human pathogen in 1982 following two outbreaks associated with consumption of
105 undercooked beef burgers in the USA[10]. It has since been reported on all continents except
106 Antarctica[11] and transmission between countries has been hypothesised to be due to transport of
107 livestock, and/or contaminated feed[12].

108

109 The current model of *E. coli* O157:H7 evolution suggests that the O serogroup conversion of an
110 ancestral Stx2-producing *E. coli* O55 to O157 and subsequent loss of β -glucuronidase activity and
111 sorbitol fermentation gave rise to the common ancestor of the current circulating population[13]. This
112 population is divided into three major lineages (I, I/II and II) with a time to common ancestor of the
113 current diversity estimated at 175 years ago[14]. Understanding the course of the global spread of *E.*
114 *coli* O157 from its common ancestor may inform action to limit its further dissemination and future
115 spread of other (emerging) foodborne pathogen including other STEC.

116

117 Analysis of whole genome sequences (WGS) of *E. coli* O157:H7 can be used to identify a core genome
118 (genes common to all strains in the analysis) and the accessory genome (genes present in at least one
119 strain, but not all)[15]. Single nucleotide polymorphisms (SNPs) can then be obtained and used to
120 generate timed phylogenies. The phylogenies obtained can be used subsequently to reconstruct a
121 detailed history of the movement of pathogens at a range of spatial scales (e.g. within and between
122 countries, as exemplified for Ebola[16] and swine flu[7]).

123

124 Here we conducted a spatio-temporal phylogenetic analysis of *E. coli* O157:H7 using the genomes of
125 757 isolates originating from four continents. Firstly, a global core-genome phylogeny was produced
126 to identify major clades followed by the reconstruction of a timed phylogeny using a representative
127 subset of isolates. Secondly, the phylodynamic analysis were visualised on a global scale and compared
128 with the history of reported cases of *E. coli* O157:H7 in different countries.

129

130 **Methods**

131

132 **Sequenced isolates.** Sequenced *E. coli* O157:H7 isolates(757) from Argentina(27), Australia(42),
133 Canada(164), the Netherlands(63), New Zealand(151), Scotland(145), Sweden(45), USA(91) and other
134 countries(29) were obtained along with associated metadata (Table S3). This sequence collection
135 comprised isolates from the following sources: human clinical(401), cattle(233), sheep(29), food and/or
136 environmental(79), and isolates of unknown origin(15). These data are available online (Table S3).

137

138 **Human *E. coli* O157:H7 incidence:** These were obtained from the national reference laboratories for
139 the above countries plus England and Wales and Japan (Table S1). A global cartogram visualising *E.*
140 *coli* O157:H7 incidence was generated utilising a geoprocessing tool[17] available from
141 <http://www.arcgis.com/home/item.html?id=d348614c97264ae19b0311019a5f2276> and implemented
142 in ArcMap 10.5.

143

144 **Analysis of genomes.**

145

146 (i) **Pan-genomic SNP analysis:** PANSEQ was used to construct a non-redundant pan-genome from all of
147 the 757 genomes[15]. This involved using a seed genome and identifying regions of 500 base pairs (bp)
148 in the seed and present in any other genome at a 99% sequence identity cut-off. Loci present in all
149 genomes underwent multiple sequence alignment and were concatenated. This aligned sequence was
150 used to identify SNPs in the core genome of all isolates (Table S7). The very high sequence identity
151 was selected to minimise the chances of recombinant regions being present in the core genome. A
152 neighbour joining tree was generated in MEGA[18].

153

154 (ii) **Bayesian phylogeographic/phylogenetic analysis:** BEAST (v1.8.2) inferred the spatiotemporal
155 dynamics of *E. coli* O157[19]. The HKY nucleotide substitution model with “Gamma+Invariant sites”,
156 with distributed rates among the sites was combined with the discrete trait substitution model (utilising
157 Bayesian Stochastic Search Variable Selection). The runs were tested for a number of population
158 models in combination with a log-normal relaxed clock for the time component and strict clock for the
159 regional component of the trees. The temporal signal of the neighbour joining tree obtained from MEGA
160 was investigated using TempEst v1.5 (<http://tree.bio.ed.ac.uk/software/tempest/>). Ancestral state

161 reconstruction was performed at the country level. Computational times meant that not all isolates could
162 be included in the analysis and a subset (n =197) were selected on the basis of representing the extent
163 of international diversity from the phylogenetic tree generated in MEGA.

164

165 The analyses were run for 100 million Markov Chain Monte Carlo (MCMC) steps and sampled every
166 2,500 steps. Convergence of parameters was checked with TRACER (v1.5), using an effective sample
167 size of 200 as the minimum to accept a model. Three independent runs were carried out to confirm
168 convergence and these were combined with the LogCombiner (v1.8.2). TreeAnnotator was used to
169 calculate the Maximum Clade Credibility Tree and the times to most recent common ancestor (MRCA)
170 with a burn-in period of 10 million MCMC states (10%). The output trees were displayed in FigTree
171 v1.4.3.

172

173 Spread (v1.0.6) and SpreaD3 (v0.9.6) were used to dynamically display the phylogeographic
174 information on Google Earth and Mozilla Firefox[20]. A major transmission event is defined as one
175 where the geographical location of the common ancestor between two nodes has changed in the
176 phylogeny. In contrast, a tip transmission event is defined as one where an isolate on the tip of the tree
177 has been isolated from a different country than its MRCA (i.e. geographical change from node to tip).

178

179 (iii) *In silico* PCR and probe-based assays were carried out both for backwards compatibility with
180 previous studies and identification of known virulence markers (See Supplementary Information
181 Section VII). This included: detection of *E. coli* O157:H7 antigen encoding, intimin and
182 enterohemolysin genes; in silico Shiga toxin subtyping; LSPA6 sub-typing, *tir* 255T and 255A
183 polymorphism analysis and Manning clade identification; SBI sub-typing and typing into Clades A-
184 G[21].

185

186 **Results**

187

188 **Recent incidence and recorded emergence of *E. coli* O157**

189 The recent incidences of *E. coli* O157:H7 infections (2010-2015) are presented in the format of a
190 cartogram illustrating the relative importance of *E. coli* O157:H7 on a country by country basis (Fig.
191 1). The first human cases of *E. coli* O157:H7 occurred in North America with a case of bloody diarrhoea
192 in the USA (California) in 1975[22] and a case in Canada in 1979, but human cases were not recorded
193 in South America (Argentina) until 1987 (Fig. 2 and Table S1). In Europe, serological evidence from
194 human cases of HUS from the Netherlands date back to 1974. However, *E. coli* O157:H7 was not
195 isolated from human cases in England and Wales until 1982, in Scotland in 1983, and in the Netherlands
196 and Sweden in 1989 (Table S1). In Australia, it was reported during 1986 - 1988 and in New Zealand

197 in 1993. Where data were available, Fig. 2 shows the increase in reported cases following the first
198 isolations.

199
200 Generally, the first isolations of *E. coli* O157:H7 from cattle occurred after the initial reports of
201 isolations in the same country from humans. An exception was Argentina, where *E. coli* O157:H7 was
202 isolated from two calves in 1977, which is generally recognised as the first cattle isolates in the
203 world[23].

204

205 **Phylogeographic emergence of *E. coli* O157**

206 A neighbour joining tree (Fig. 3), containing all of the genomes from this study, readily demonstrates
207 that representatives from individual countries exhibit distinct clustering. Genomes are clustered around
208 the tree in seven clades labelled A to G. The relationship of these clades with previous DNA based
209 typing systems for *E. coli* O157:H7 (provided in the supplementary material section VIII). Only the
210 Netherlands is represented in all clades, while the USA is represented in all clades except Clade C.
211 Argentina is present in the fewest clades (2), followed by New Zealand and Canada (3), and Australia
212 (4). The tree was generated from 3956 SNPs obtained from the core genome (730kb). Both animal and
213 human clinical isolates were dispersed across the tree except for Clade A which comprised eight clinical
214 and two isolates from unknown sources (Fig. 3).

215

216 TempEst demonstrated a poor correlation of genetic divergence through time. Hence, a relaxed, un-
217 correlated, log normally distributed clock (UCLD), as used previously for influenza A viruses in
218 swine[7], was applied in BEAST, which enabled each branch of the tree to have its own evolutionary
219 rate. The exponential growth and birth-death population models both converged and a Bayes factor
220 (BF) of 4.28 provides positive support of the fit of the exponential growth model[24].

221

222 The common ancestor of the isolates in this study using the exponential growth model was predicted
223 (with a probability of 0.66) to have originated in the Netherlands around 1890 (Bayesian 95% credible
224 interval 1845–1925) (Fig. 4). The birth-death model also found the Netherlands as the common ancestor
225 around 1910 (1886-1932). The exponential growth model gives the second most probable ancestral
226 origin as Scotland (probability 0.19), suggesting that Europe is the most likely origin for these isolates.
227 The USA, Canada and Australia all showed relatively low probabilities of being the ancestral origin
228 (0.06, 0.04 and 0.02, respectively). Clades A to E were predicted to have had a Dutch common ancestor
229 from around 1905 (1870–1937), similar to the predicted date 1910 (1875–1940) for the separate
230 common Dutch ancestor of Clades F and G.

231

232 The analysis predicts 34 main transmission events between countries (Table S2) of which 21 and 13
233 were intra- and inter-continental, respectively. The earliest country to country transmissions were all
234 intercontinental between the Netherlands and the following countries: Clade D to Australia in 1937
235 (1909–1958), Clade G to the USA in 1941 (1927–1966) and Clade E(ii) to the USA in 1949 (1927–
236 1966) (Fig. 5 and Figures S1 and S2). Figure 5 shows that by 1985, less than ten years after the first
237 reported case of *E. coli* O157:H7 in humans, the organism was present on at least four continents. The
238 emergence of *E. coli* O157:H7 can be described in detail by individual country or by clade and this is
239 provided in the supplementary material (Sections IV and V).

240

241 **Identification of virulent clades**

242 Virulent clades comprise *E. coli* O157:H7 isolates that have the ability to cause the most severe clinical
243 disease (e.g. severe or bloody diarrhoea and HUS). There have been several reports indicating that *stx*₂
244 and in particular *stx*_{2a} and *stx*_{2d} positive isolates exhibit greater morbidity than *stx*₁ and other *stx*₂
245 isolates[25]. Of the 757 genomes in the present study, 476 (62.9%) harbour *stx*_{2a} and none carry *stx*_{2d}
246 (Fig. 6). Clades F and G contain the highest proportion (>75%) of *stx*_{2a} positive strains. In contrast,
247 Clades B, C, D and E, that dominate the lower branch of the BEAST phylogeny (Fig. 4), have very low
248 carriage (<7%) of *stx*_{2a}. This suggests that *stx*_{2a} was not a characteristic of the common ancestor in this
249 part of the phylogeny, but likely was introduced by *stx*_{2a} carrying phage in more recent times. This can
250 also be visualised by the distribution of SBI types in Fig. S1g.

251

252 **Discussion**

253 This study provides the first comprehensive global phylogeographical analysis of Shiga toxin-
254 producing *E. coli* O157:H7. The common ancestor of the current circulating diversity was estimated to
255 have originated in the Netherlands (i.e. mainland Europe) around 1890 (1845–1925). This timeframe is
256 very similar to a recent UK study on UK genomes, of around 1840 (1817–1855)[26]. Although the
257 earliest reported cases of human disease are from North America[22], a retrospective examination of
258 sera from patients with HUS suggests an early presence (mid 1970s) of *E. coli* O157:H7 in the
259 Netherlands[27]. However, the first reporting date of isolates from diseased humans within a region is
260 dependent on several factors including: the presence of *E. coli* O157:H7 in a geographical area; its
261 virulence (severity of clinical symptoms), the availability of detection methods, and the
262 expertise/awareness of public health officials.

263
264 Similar phylogeographic studies on infectious disease transmission have been performed (e.g.
265 Ebola[16]) where person-to-person transmission is a key mechanism of spread. In contrast, *E. coli*
266 O157:H7 has limited person-to-person transmission (most cases considered sporadic and only 10 -15%
267 outbreak cases are secondary transmission[28]) and the global spread is more likely to be related to its
268 epidemiology in animals. Long range animal movements have been found to play an important role in
269 the global migration of swine flu[7]. There are several other ways *E. coli* O157:H7 can potentially be
270 spread across large geographical distances[12]. First, contaminated animal feed has been reported for
271 feedlot cattle in the mid-western USA[29] and feed can be exported or imported from overseas
272 (<http://www.food.gov.uk>). Second, wild animals including birds can be relevant at regional or sub-
273 continental scales[30]. However, bird migration is unlikely to explain the longitudinal transmission
274 routes since most bird migration occurs latitudinally. Third, international movement of other farm
275 animals including pigs, goats and turkeys, which occasionally shed this pathogen[30]. Fourth, the global
276 trade and transportation of contaminated food[26]. Since the chains of human infection are usually short
277 it is more likely that this mechanism will result in tip transmissions. Altogether, animal movements, of
278 cattle and sheep, are considered the most likely transmission pathway for *E. coli* O157:H7 to establish
279 a long term presence in a country or region.

280
281 The Netherlands being the origin of the current diversity of *E. coli* O157:H7 is plausible, as the country
282 extensively exported Holstein-Friesian cattle across the world[31]. This German-Dutch breed, known
283 for its high milk production, has been successfully introduced across the world where the climate and
284 conditions are suited for European cattle, including North America (1850s), South America (1880s) and
285 Japan (>100 years ago). There has been a long history of cattle movement between the USA and Canada,
286 with approximately 500,000 animals crossing the border each year during the 1980s, which is
287 contemporaneous with the predicted transmission events between these countries

288 (<https://www.usitc.gov/publications/332/pub2591.pdf>). Similarly, cattle and sheep imports into New
289 Zealand from Australia and the UK have occurred since the 1860s and throughout the 20th century[32].
290 We were not able to quantitatively link the global spread of *E. coli* O157:H7 to the global cattle trade
291 in the 19th and early 20th century as these historic data are incomplete but the importance of massive
292 trans-atlantic cattle movements in the emergence and spread of O157 has been postulated
293 previously[33].

294
295 The phylodynamic predictions infer the likely dates of introduction of *E. coli* O157:H7 into the
296 countries under study were either before or about the same time as the first isolates were obtained from
297 humans and/or cattle (Fig. 1 and Fig. 7). The Netherlands and Australia were the only countries where
298 *E. coli* O157:H7 was predicted to be present >50 years before human case reports. However, for
299 Australia, the first transmission wave (Clade D) carried generally the less potent Stx2c form of the Shiga
300 toxin. Only with the second wave when *stx*_{2a} Clade G strains were introduced around 1994, did it
301 become likely that severe disease would occur, resulting in a greater need to investigate the aetiology
302 of those cases. Since the Netherlands is predicted to harbour the common ancestor (1910 (1875–1940))
303 for the more virulent Clades F and G, which contain the *stx*_{2a} gene, it would be expected that cases of
304 severe disease associated with these pathogens would occur from this date onwards. Unfortunately,
305 serological evidence only dates back to 1974[27], but it is likely that there were human cases prior to
306 this. Alternatively, the prevalence in the cattle population may have been low at this time, resulting in
307 minor or negligible disease rates in the human population.

308
309 The main limitation of the present study is lack of representative genomes from a number of countries
310 where *E. coli* O157:H7 is known to be present (e.g. elsewhere in Europe, Japan, China, Brazil, and the
311 under-represented countries of the African continent). As a result, inferences in disease transmission in
312 this paper must be considered in the context of such missing data, which may involve movement through
313 intermediate countries[7]. The Netherlands being predicted as the origin of the current circulating
314 diversity should be treated with some caution as it in fact may act as a proxy for central Europe. Finally,
315 only a representative sub-set of all isolates could be used in the BEAST analysis due to computational
316 requirements.

317
318 Knowledge of the global spread of *E. coli* O157:H7 enables further insights into how to mitigate the
319 effects of this pathogen and reduce the risk of future STEC emergence [34]. This study informs on how
320 non-O157 STEC may spread globally in similar fashion as *E. coli* O157. For biosecurity and trade
321 purposes, movements of live animals are now recorded between a number of countries
322 (<http://comtrade.un.org>) and it may be sensible to test animals for STEC prior to transport, as well as
323 animal feed for any STEC prior to transportation/shipment. This would be particularly important when
324 a new virulent strain of STEC has been detected in a country to prevent its further spread. General test

325 methods for detection and sequencing of *E. coli* O157:H7 and other STEC are available[35, 36]. and
326 international collaboration will be critical here. Furthermore, computer simulations of disease
327 emergence and transmission would help identify countries at higher risk, and inform where surveillance
328 and control strategies for animal movements should take place[7]. Additionally, promotion of trade in
329 germplasm would also lessen the chance of transmission.

330 In conclusion, if the measures mentioned above are not carried out, it is likely that new STEC strains
331 will emerge and spread around the world and future generations will continue to suffer disease from
332 this group of bacterial pathogens.

333

334 **Acknowledgements**

335

336 The authors acknowledge the support of Food Standards Scotland (FS102029) and the University of
337 Aberdeen for funding sequencing of the Scottish *E. coli* O157:H7 genomes. New Zealand clinical
338 isolates were provided by the Institute of Environmental Science and Research, funded by the Ministry
339 of Health. The Public Health Agency of Canada for the sequencing of Canadian *E. coli* O157:H7
340 genomes. We would like to thank the Canadian National Enteric Surveillance Program (NESP) of the
341 Public Health Agency of Canada for providing national incidence rates of *E. coli* O157:H7 infection in
342 Canada, specifically Frank Pollari and Kimberly Nguy from the Centre for Food-borne, Environmental
343 and Zoonotic Infectious Diseases, and the Enteric Diseases Program of the National Microbiology
344 Laboratory; we would also like to thank the provincial public health laboratories for providing this
345 information to the NESP. We would also like to thank Associate Prof Adrian Cookson for the
346 sequencing of New Zealand *E. coli* O157:H7 genomes and epidemiological data for human cases, help
347 and advice with the New Zealand isolates and Springer Browne from Massey University for providing
348 New Zealand cattle import data. We acknowledge the Public Health Microbiology Laboratory,
349 Queensland Department of Health, for the provision of Queensland human isolates. The mention of a
350 trade name, proprietary product, or specific equipment does not constitute a guarantee or warranty by
351 the USDA and does not imply approval to the exclusion of other products that might be suitable. The
352 USDA is an equal opportunity employer and provider.

353

354

355 **Funding**

356 This work was supported by: Scotland by Food Standards Scotland [Grant Number FS102029]
357 and University of Aberdeen; New Zealand, Institute of Environmental Science and Research;
358 Canada, the Public Health Agency of Canada; United States, United States Department of
359 Agriculture.

360

361 **Author contributions**

362 EF, NS, KG, JLB, KF and FPR designed the research

363 EF, NS and OR wrote the manuscript.

364 EF, OR, BL, MM, JLB, CL, VG, RS, AH, IF, NF, TG, PB, PJ, MR, IC, JC, CJ, KG, GM, PSC,

365 FPR, KF and NS generated and interpreted the data used in the analysis

366 OR, NS, CL and FPR performed the analysis

367 All authors reviewed the manuscript

368

369 **Conflict of interest**

370 NS reports personal fees from Food Standards Scotland in his role as Chief Scientific Advisor during
371 the conduct of the study.

372

373 **Disclaimer**

374 The opinions expressed in this paper are the authors own and do not reflect the view of any of
375 the organisations that they work for.

376

377

378 **LEGENDS**

379 Figure 1. Cartogram of the incidence of *E. coli* O157:H7 per 10 million (per year) from 2010 – 2015
380 where the area of the country corresponds to the incidence (Note Australia incidence is based on the
381 years 2001-2009 which were the only available data), and insert is a map of the world of original scale
382 with countries coloured black where data were available.

383
384 Figure 2. The reported human cases of *E. coli* O157:H7 by country obtained from national reference
385 laboratories (red arrow indicates first human case, green arrow first human case of HUS associated with
386 *E. coli* O157:H7 and blue arrow first isolation from cattle). The horizontal coloured bars and filled dots
387 represent the 95% credible intervals and most likely date of the first major introduction estimated by
388 the BEAST analysis (there is no bar and dot for England/Wales and only the dot for Japan as there were
389 0 and 1 genomes only from these countries in the current study).

390
391 Figure 3. Nearest-neighbour joining tree of 757 *E. coli* O157:H7 isolates inferred from the 3956 SNP's
392 obtained from PANSEQ: Scotland (●); Canada (●); USA (●); the Netherlands (●); Sweden (●); New
393 Zealand (●); Australia (●); Europe (●); Italy (●); Egypt (○); Asia (△); Argentina (▲); South America
394 (△); Unknown (●). Letters indicate branches of Clades A to G. The scale marker indicates genetic
395 distance in SNPs. The pie charts indicate proportion of a particular source in a clade (human clinical –
396 (■); animal -(□); and other (food, environmental and unknown) -(≡)). The location of the root of the
397 tree using an *E. coli* O55:H7 (strain CB9615) is highlighted [21].).

398
399 Figure 4. Bayesian Maximum Clade Credibility (MCC) phylogeographic tree for 197 *E. coli* O157:H7
400 isolates visualised by FigTree. Branch colours correspond to the most probable ancestral geographic
401 location. Clades A–G are marked on the phylogeny. The dates of the transmission events are listed in
402 Supplementary Table S2. The lower figure provides a demographic reconstruction of the population
403 size using exponential growth rate.

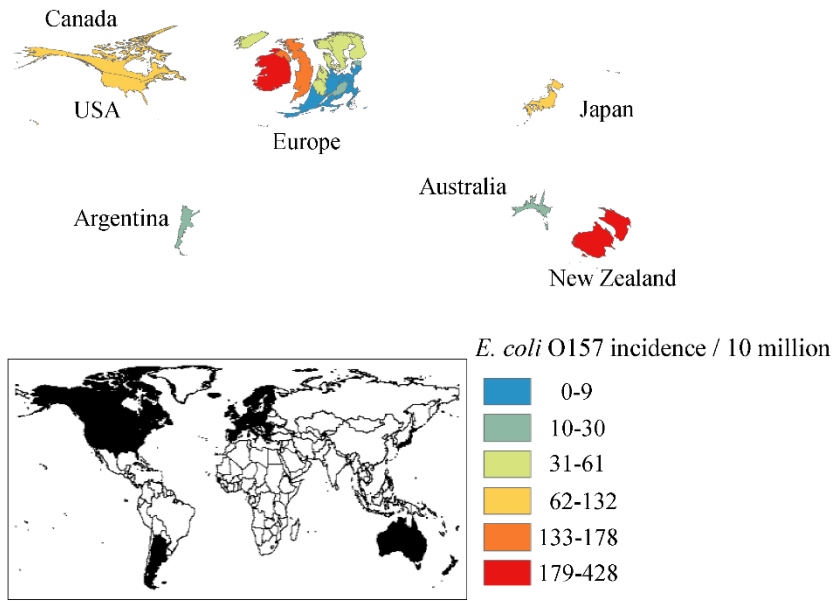
404
405 Figure 5. Geographic dynamics of the time to transmission of *E. coli* O157:H7. The arrows indicate 34
406 major transmission events and dates are the median values of the MRCA taken from BEAST. The letters
407 denote the phylogenetic clades (see Fig. 3). The map was based on the output from Spread3 and
408 reconstructed in ArcMap 10.5. The map can be viewed dynamically in Google Earth using the kml file
409 (Supplementary File S1.kml) or by video (Supplementary File S2.wmv see Fig. 7).

410
411 Figure 6. Frequency of *E. coli* O157:H7 Shiga toxin genes by clade for (a) *stx*_{1a}, (b) *stx*_{2a}, (c) *stx*_{2c} and
412 (d) *stx* negative.

413
414 Figure 7. Screenshot of video illustrating the global spread of *E. coli* O157:H7
415 (Supplementary_File_S2.mp4).

416

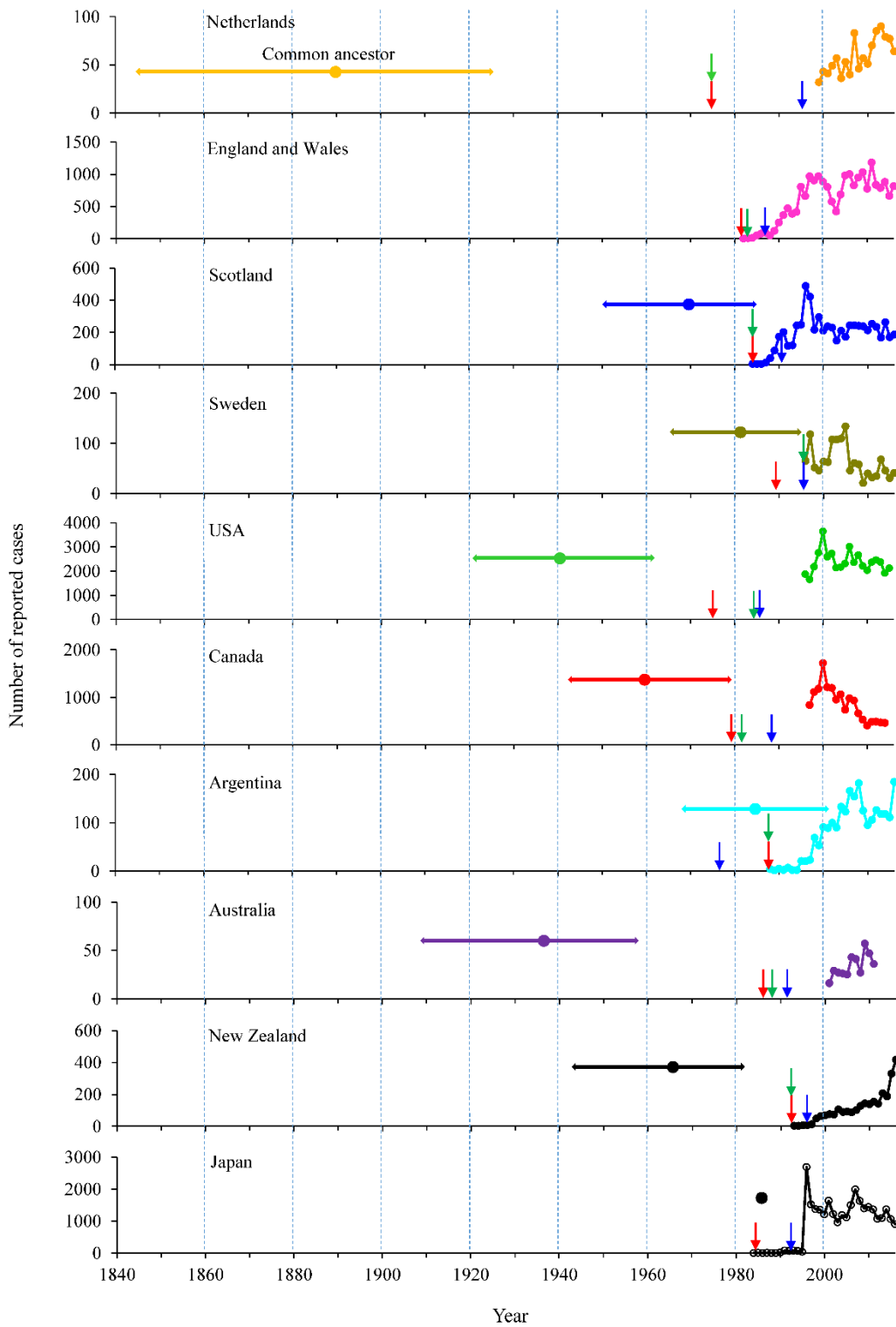
417



418

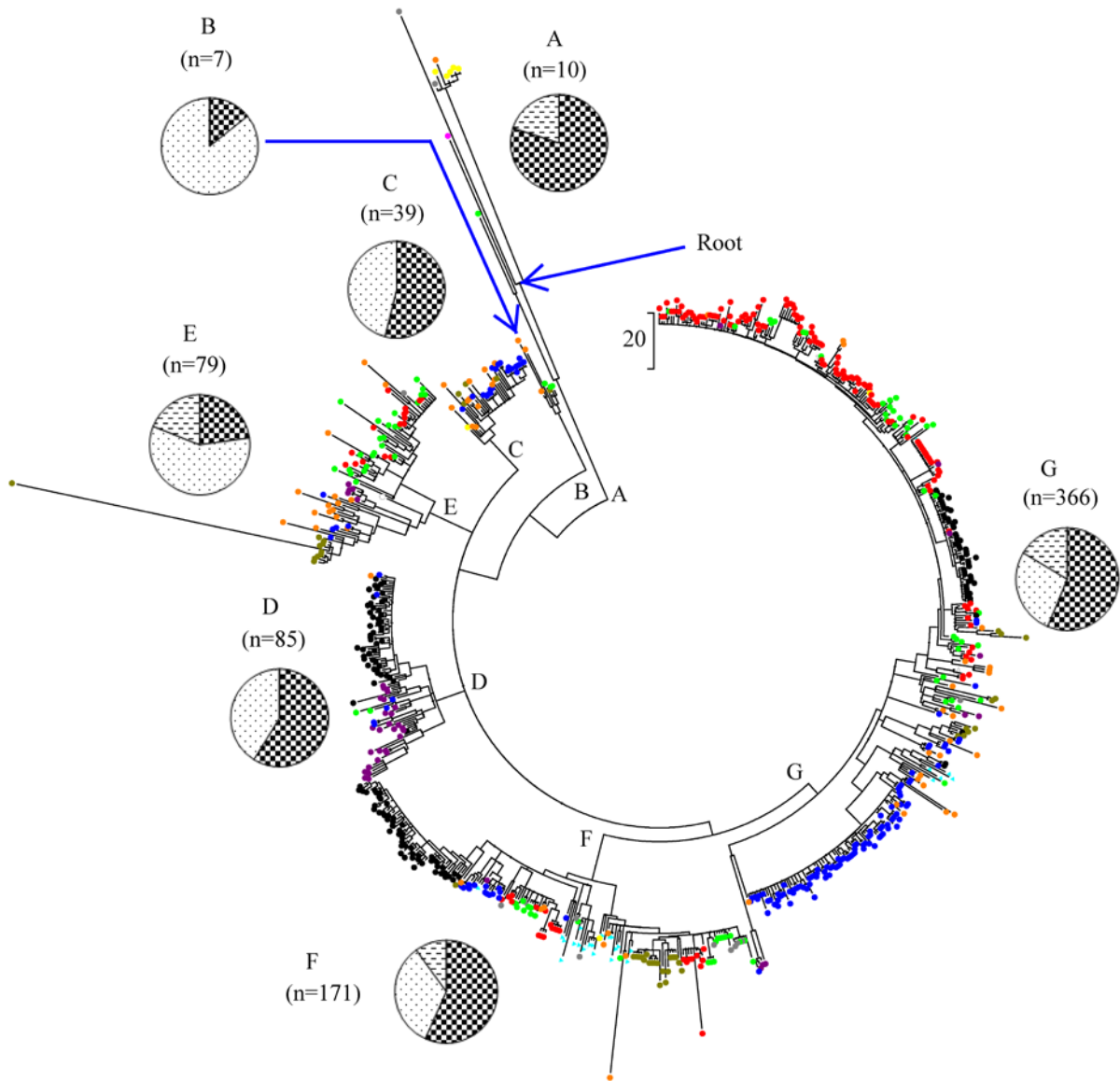
419 Figure 1. Cartogram of the incidence per 10 million (per year) of *E. coli* O157:H7 from 2010 – 2015
420 where the area of the country corresponds to the incidence (Note Australia incidence is based on the
421 years 2001-2009 which were the only available data), and inset is a map of the world of original scale
422 with countries coloured black where data were available.

423



425

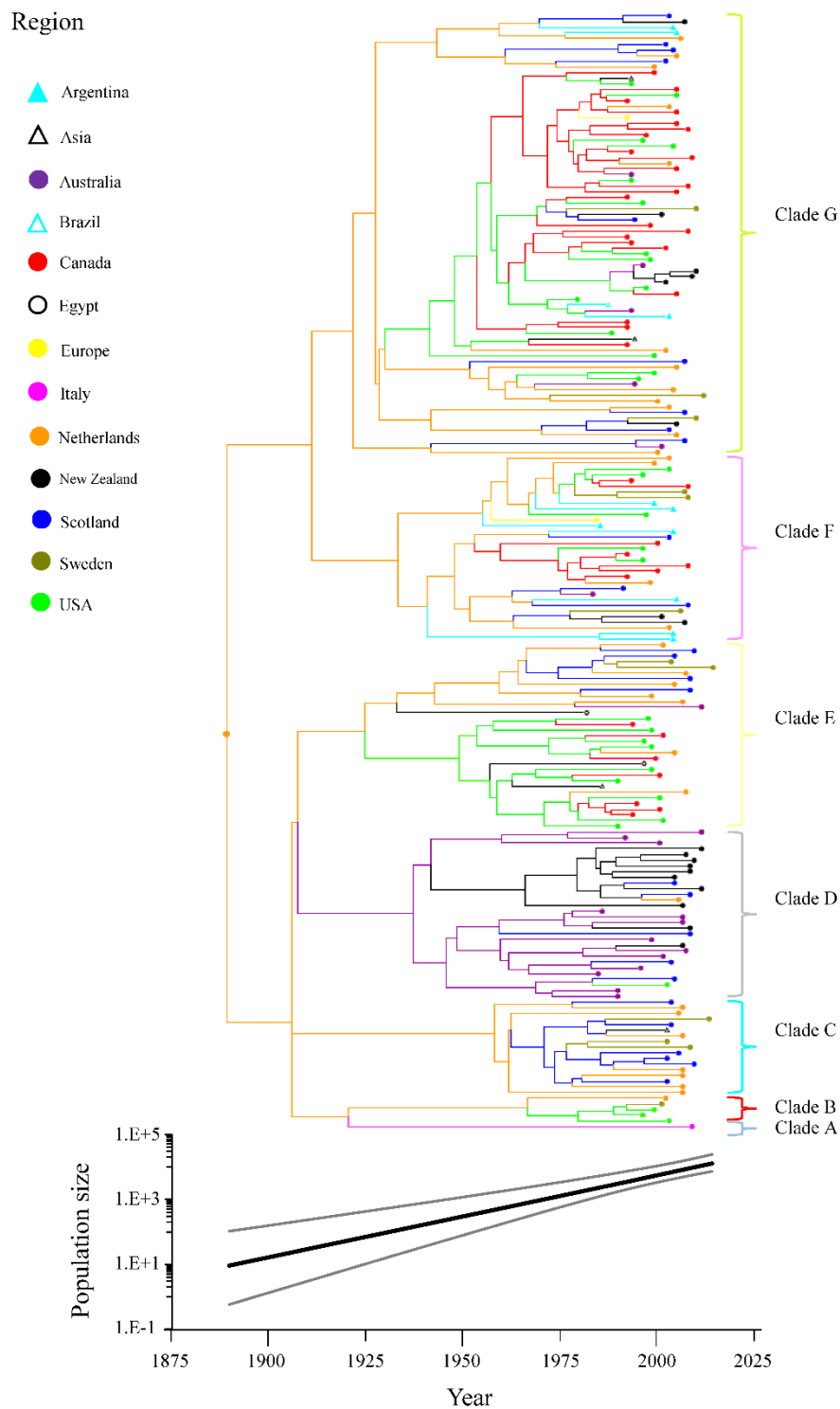
426 Figure 2. The reported human cases of *E. coli* O157:H7 by country obtained from national reference
 427 laboratories (red arrow indicates first human case, green arrow first human case of HUS associated with
 428 *E. coli* O157:H7 and blue arrow first isolation from cattle). The horizontal coloured bars and filled dots
 429 represent the 95% credible intervals and most likely date of the first major introduction estimated by
 430 the BEAST analysis (there is no bar and dot for England/Wales and only the dot for Japan as there were
 431 0 and 1 genomes only from these countries in the current study).



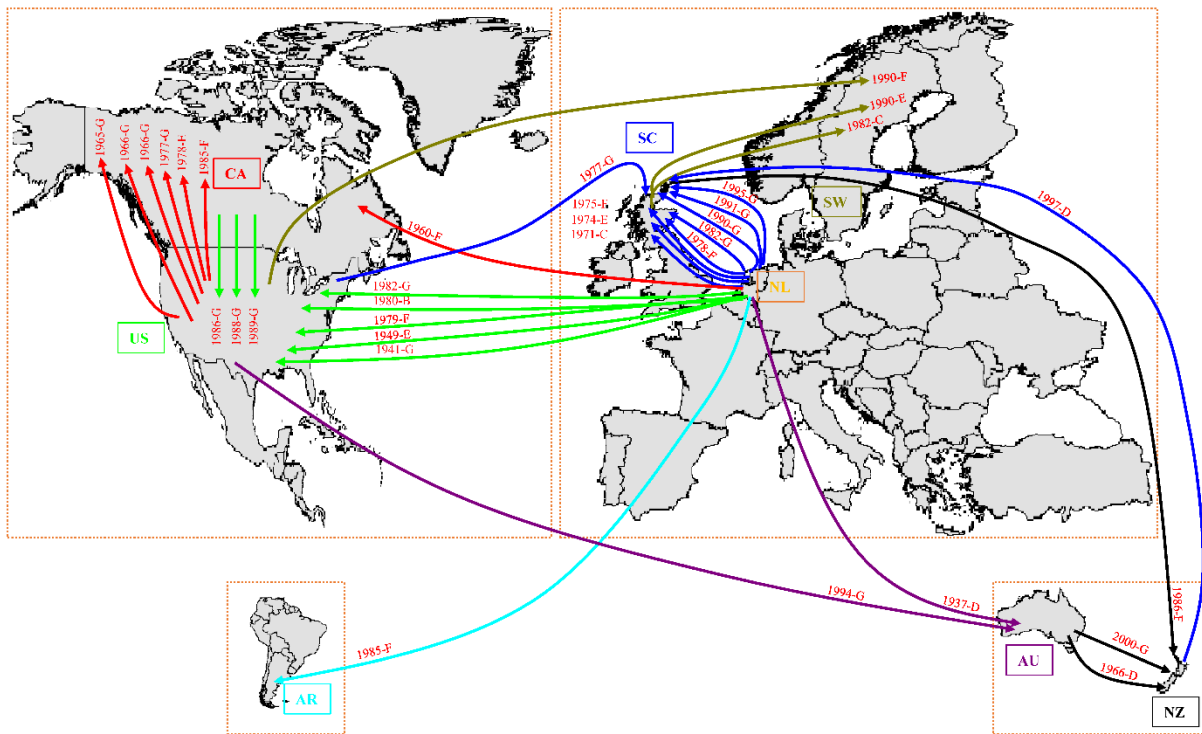
432
433

434 Figure 3. Nearest-neighbour joining tree of 757 *E. coli* O157:H7 isolates inferred from the 3956 SNP's
 435 obtained from PANSEQ: Scotland (●); Canada (●); USA (●); the Netherlands (●); Sweden (●); New
 436 Zealand (●); Australia (●); Europe (●); Italy (●); Egypt (○); Asia (△); Argentina (▲); South America
 437 (▲); Unknown (●). Letters indicate branches of Clades A to G. The scale marker indicates genetic
 438 distance in SNPs. The pie charts indicate proportion of a particular source in a clade (human clinical –
 439 (▣); animal -(□); and other (food, environmental and unknown) -(≡)). The location of the root of the
 440 tree using an *E. coli* O55:H7 (strain CB9615) is highlighted [21].

441

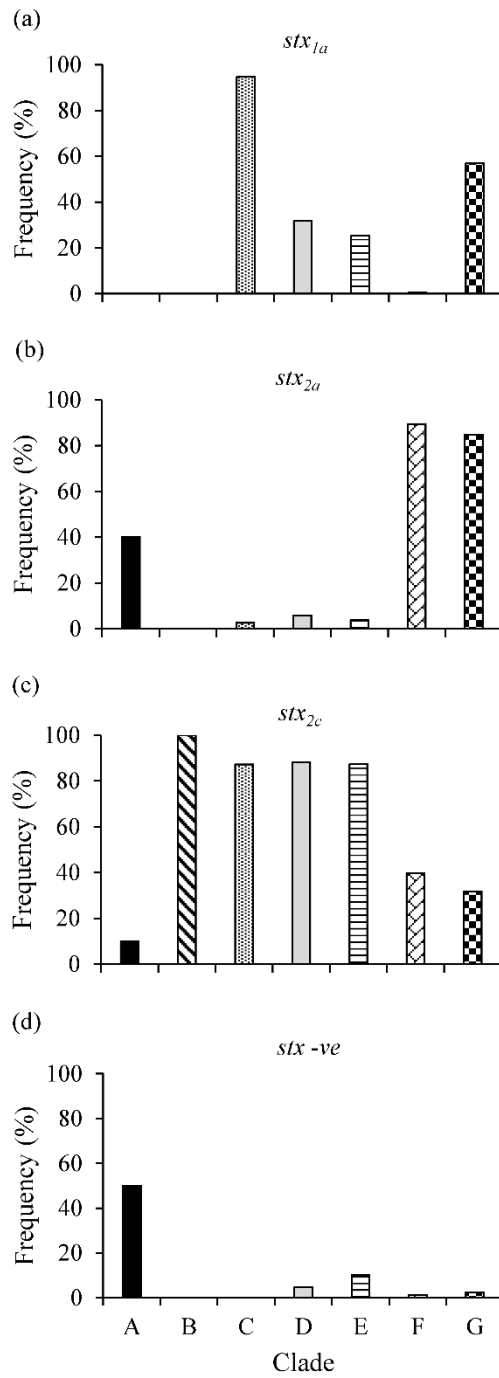


442
 443 Figure 4. Bayesian Maximum Clade Credibility (MCC) phylogeographic tree for 197 *E. coli* O157:H7
 444 isolates visualised by FigTree. Branch colours correspond to the most probable ancestral geographic
 445 location. Clades A–G are marked on the phylogeny. The dates of the transmission events are listed in
 446 Supplementary Table S2. The lower figure provides a demographic reconstruction of the population
 447 size using exponential growth rate.
 448



449
 450
 451
 452
 453
 454
 455
 456

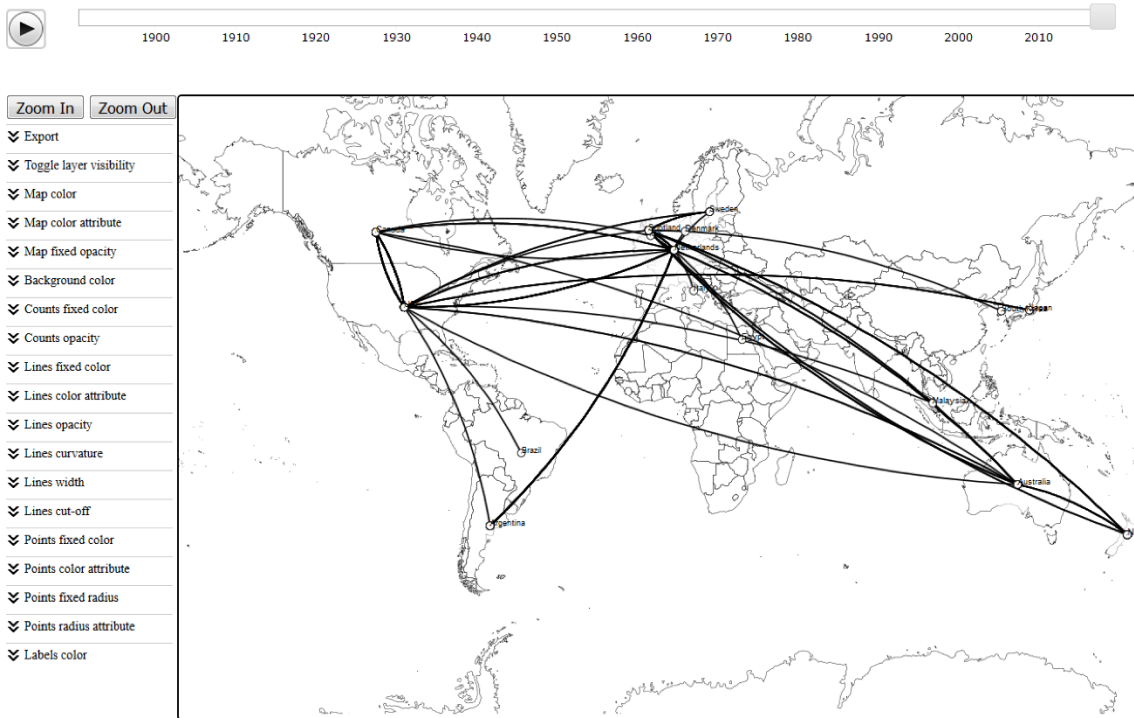
Figure 5. Geographic dynamics of the transmission of *E. coli* O157:H7. The arrows indicate 34 major transmission events and dates are the median values of the MRCA taken from BEAST. The letters denote the phylogenetic clades (see Fig. 3). The map was based on the output from SpreadD3 and reconstructed in ArcMap 10.5. The map can be viewed dynamically in Google Earth using the kml file (Supplementary File S1.kml) or by video (Supplementary File S2.wmv and Fig. 7).



457

458 Figure 6. Frequency of *E. coli* O157:H7 Shiga toxin genes by clade for (a) *stx*_{1a}, (b) *stx*_{2a}, (c) *stx*_{2c} and
 459 (d) *stx* negative.

460



461
 462
 463
 464
 465

Figure 7. Screenshot of video illustrating the global spread of *E. coli* O157:H7 (Supplementary_File_S2.mp4).

- 467 1. Morens DM, Fauci AS. Emerging infectious diseases in 2012: 20 years after the institute of
468 medicine report. *Mbio*, **2012**; 3: e00494-12.
- 469 2. Jones KE, Patel NG, Levy MA, et al. Global trends in emerging infectious diseases. *Nature*, **2008**;
470 451: 990-U4.
- 471 3. Weill F, Domman D, Njamkepo E, et al. Genomic history of the seventh pandemic of cholera in
472 africa. *Science*, **2017**; 358: 785.
- 473 4. Holmes AH, Moore LSP, Sundsfjord A, et al. Understanding the mechanisms and drivers of
474 antimicrobial resistance. *Lancet*, **2016**; 387: 176-87.
- 475 5. Jones BA, Grace D, Kock R, et al. Zoonosis emergence linked to agricultural intensification and
476 environmental change. *Proc Natl Acad Sci U S A*, **2013**; 110: 8399-404.
- 477 6. Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP, Daszak P. Predicting the
478 global spread of H5N1 avian influenza. *Proc Natl Acad Sci U S A*, **2006**; 103: 19368-73.
- 479 7. Nelson MI, Schaefer R, Gava D, Cantao ME, Ciacci-Zanella JR. Influenza A viruses of human
480 origin in swine, Brazil. *Emerg Infect Dis*, **2015**; 21: 1339-47.
- 481 8. Karmali MA, Gannon V, Sargeant JM. Verocytotoxin-producing *Escherichia coli* (VTEC). *Vet*
482 *Microbiol*, **2010**; 140: 360-70.
- 483 9. Mughini Gras L, van Pelt W, van der Voort M, Heck M, Friesema I, Franz E. Attribution of human
484 infections with shiga toxin-producing *Escherichia coli* (STEC) to livestock sources and identification
485 of source-specific risk factors, the Netherlands (2010–2014). *Zoonoses and Public Health*, **2017**.
- 486 10. Riley L, Remis R, Helgerson S, et al. Hemorrhagic colitis associated with a rare *Escherichia coli*
487 serotype. *N Engl J Med*, **1983**; 308: 681-5.
- 488 11. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. Super-shedding and the link
489 between human infection and livestock carriage of *Escherichia coli* O157. *Nat Rev Microbiol*, **2008**;
490 6: 904-12.
- 491 12. Davis M, Hancock D, Besser T, et al. Correlation between geographic distance and genetic
492 similarity in an international collection of bovine faecal *Escherichia coli* O157 : H7 isolates.
493 *Epidemiol Infect*, **2003**; 131: 923-30.
- 494 13. Feng P, Lampel K, Karch H, Whittam T. Genotypic and phenotypic changes in the emergence of
495 *Escherichia coli* O157 : H7. *J Infect Dis*, **1998**; 177: 1750-3.
- 496 14. Dallman TJ, Ashton PM, Byrne L, et al. Applying phylogenomics to understand the emergence of
497 shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK.
498 *Microb Genom*, **2015**; 1: e000029.
- 499 15. Laing C, Buchanan C, Taboada EN, et al. Pan-genome sequence analysis using panseq: An online
500 tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics*, **2010**; 11:
501 461,2105-11-461.

- 502 16. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained
503 the ebola epidemic. *Nature*, **2017**.
- 504 17. Gastner M, Newman M. Diffusion-based method for producing density-equalizing maps. *Proc*
505 *Natl Acad Sci U S A*, **2004**; 101: 7499-504.
- 506 18. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary
507 analysis of DNA and protein sequences. *Brief Bioinform*, **2008**; 9: 299-306.
- 508 19. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the
509 BEAST 1.7. *Mol Biol Evol*, **2012**; 29: 1969-73.
- 510 20. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. Spread3: Interactive
511 visualization of spatiotemporal history and trait evolutionary processes. *Mol Biol Evol*, **2016**; 33:
512 2167-9.
- 513 21. Strachan NJC, Rotariu O, Lopes B, et al. Whole genome sequencing demonstrates that geographic
514 variation of *Escherichia coli* O157 genotypes dominates host association. *Scientific Reports*, **2015**; 5:
515 14145.
- 516 22. Wells J, Davis B, Wachsmuth I, et al. Laboratory investigation of hemorrhagic colitis outbreaks
517 associated with a rare *Escherichia coli* serotype. *J Clin Microbiol*, **1983**; 18: 512-20.
- 518 23. Orskov F, Orskov I, Villar J. Cattle as reservoir of verotoxin-producing *Escherichia coli*
519 O157:H7. *Lancet*, **1987**; 2: 276-.
- 520 24. Drummond AJ, Bouckaert RR. Bayesian evolutionary analysis with BEAST. 1st ed. Cambridge,
521 United Kingdom: Cambridge University Press, **2015**.
- 522 25. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. Shiga toxin subtypes display dramatic
523 differences in potency. *Infect Immun*, **2011**; 79: 1329-37.
- 524 26. Dallman TJ, Byrne L, Ashton PM, et al. Whole-genome sequencing for national surveillance of
525 shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis*, **2015**; 61: 305-12.
- 526 27. Chart H, Rowe B, vd Kar N, Monnens LA. Serological identification of *Escherichia coli* O157 as
527 cause of haemolytic uraemic syndrome in netherlands. *Lancet*, **1991**; 337: 437.
- 528 28. Snedeker KG, Shaw DJ, Locking ME, Prescott RJ. Primary and secondary cases in *Escherichia*
529 *coli* O157 outbreaks: A statistical analysis. *BMC Infect Dis*, **2009**; 9: 144.
- 530 29. Dodd CC, Sanderson MW, Sargeant JM, et al. Prevalence of *Escherichia coli* O157 in cattle feeds
531 in midwestern feedlots. *Appl Environ Microbiol*, **2003**; 69: 5243-7.
- 532 30. Ferens WA, Hovde CJ. *Escherichia coli* O157:H7: Animal reservoir and sources of human
533 infection. *Foodborne Pathog Dis*, **2011**; 8: 465-87.
- 534 31. Houghton FL. Holstein-friesian cattle: A history of the breed and its development in america. New
535 York, USA: Cornell University Library, **1897**.
- 536 32. Binney BM, Biggs PJ, Carter PE, Holland BM, French NP. Quantification of historical livestock
537 importation into New Zealand 1860-1979. *N Z Vet J*, **2014**; 62: 309-14.

- 538 33. Leopold SR, Magrini V, Holt NJ, et al. A precise reconstruction of the emergence and constrained
539 radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad
540 Sci U S A, **2009**; 106: 8713-8.
- 541 34. Tozzoli R, Grande L, Michelacci V, et al. Shiga toxin-converting phages and the emergence of
542 new pathogenic *Escherichia coli*: A world in motion. Front Cell Infect Microbiol, **2014**; 4: 80.
- 543 35. Franz E, Delaquis P, Morabito S, et al. Exploiting the explosion of information associated with
544 whole genome sequencing to tackle shiga toxin-producing *Escherichia coli* (STEC) in global food
545 production systems (vol 187, pg 57, 2014). Int J Food Microbiol, **2015**; 193: 159-.
- 546 36. Nadon C, Van Walle I, Gerner-Smidt P, et al. PulseNet international: Vision for the
547 implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro
548 Surveill, **2017**; 22: 10.2807/1560,7917.ES.2017.22.23.30544.
- 549