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

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

AUTHORS

Eelco Franz[1], Ovidiu Rotariu[2], Bruno S. Lopes[3], Marion MacRae[3], James L. Bono[4], Chad Laing[5], Victor Gannon[5], Robert Söderlund[6], Angela H.A.M. van Hoek[1], Ingrid Friesema[1], Nigel P. French[7], Tessy George[7], Patrick J. Biggs[7], Patricia Jaros[7], Marta Rivas[8], Isabel Chinen[8], Josefina Campos[8], Cecilia Jernberg[9], Kari Gobius[10], Glen E. Mellor[10], P. Scott Chandry[10], Francisco Perez-Reche[11], Ken J. Forbes[3] and Norval J.C. Strachan[2*]

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

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

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

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

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

[6]National Veterinary Institute (SVA), Uppsala, Sweden.

[7] [m]EpiLab, Infectious Disease Research Centre, School of Veterinary Science, Massey University, Palmerston North, New Zealand.

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

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

[10]CSIRO Agriculture and Food, Werribee, Australia.

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

Keywords:

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

[*]**Corresponding author**:

Name: Prof. Norval Strachan

Address: School of Biological Sciences, The University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen, United Kingdom, AB24 3UU

Tel no: +44 1224 272699

Fax no: +44 1224 272703

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

Alternate Corresponding Author:

Name: Dr. Eelco Franz

Address: Head of Department Epidemiology and Surveillance of Enteric and Zoonotic Infections, Centre for Infectious Disease Control (CIb), National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, the Netherlands

Tel.: +31 30 2747063

Fax: +31 6 15502479

Email address: eelco.franz@rivm.nl

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

Summary

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

Abstract

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

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

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

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

Introduction

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

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

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

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

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

Methods

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

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

Analysis of genomes.

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

(ii) *Bayesian phylogeographic/phylodynamic analysis:* BEAST (v1.8.2) inferred the spatiotemporal dynamics of *E. coli* O157[19]. The HKY nucleotide substitution model with "Gamma+Invariant sites", with distributed rates among the sites was combined with the discrete trait substitution model (utilising Bayesian Stochastic Search Variable Selection). The runs were tested for a number of population models in combination with a log-normal relaxed clock for the time component and strict clock for the regional component of the trees. The temporal signal of the neighbour joining tree obtained from MEGA was investigated using TempEst v1.5 (http://tree.bio.ed.ac.uk/software/tempest/). Ancestral state reconstruction was performed at the country level. Computational times meant that not all isolates could be included in the analysis and a subset (n =197) were selected on the basis of representing the extent of international diversity from the phylogenetic tree generated in MEGA.

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

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

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

Results

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

The recent incidences of *E. coli* O157:H7 infections (2010-2015) are presented in the format of a cartogram illustrating the relative importance of *E. coli* O157:H7 on a country by country basis (Fig. 1). The first human cases of *E. coli* O157:H7 occurred in North America with a case of bloody diarrohea in the USA (California) in 1975[22] and a case in Canada in 1979, but human cases were not recorded in South America (Argentina) until 1987 (Fig. 2 and Table S1). In Europe, serological evidence from human cases of HUS from the Netherlands date back to 1974. However, *E. coli* O157:H7 was not isolated from human cases in England and Wales until 1982, in Scotland in 1983, and in the Netherlands and Sweden in 1989 (Table S1). In Australia, it was reported during 1986 - 1988 and in New Zealand in 1993. Where data were available, Fig. 2 shows the increase in reported cases following the first isolations.

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

Phylogeographic emergence of *E. coli* O157

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

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

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

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

Identification of virulent clades

Virulent clades comprise *E. coli* O157:H7 isolates that have the ability to cause the most severe clinical disease (e.g. severe or bloody diarrhoea and HUS). There have been several reports indicating that stx_2 and in particular stx_{2a} and stx_{2d} positive isolates exhibit greater morbidity than stx_1 and other stx_2 isolates[25]. Of the 757 genomes in the present study, 476 (62.9%) harbour stx_{2a} and none carry stx_{2d} (Fig. 6). Clades F and G contain the highest proportion (>75%) of stx_{2a} positive strains. In contrast, Clades B, C, D and E, that dominate the lower branch of the BEAST phylogeny (Fig. 4), have very low carriage (<7%) of stx_{2a}. This suggests that stx_{2a} was not a characteristic of the common ancestor in this part of the phylogeny, but likely was introduced by stx_{2a} carrying phage in more recent times. This can also be visualised by the distribution of SBI types in Fig. S1g.

Discussion

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

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

The Netherlands being the origin of the current diversity of *E. coli* O157:H7 is plausible, as the country extensively exported Holstein-Friesian cattle across the world[31]. This German-Dutch breed, known for its high milk production, has been successfully introduced across the world where the climate and conditions are suited for European cattle, including North America (1850s), South America (1880s) and Japan (>100 years ago). There has been a long history of cattle movement between the USA and Canada, with approximately 500,000 animals crossing the border each year during the 1980s, which is contemporaneous with the predicted transmission events between these countries (https://www.usitc.gov/publications/332/pub2591.pdf). Similarly, cattle and sheep imports into New Zealand from Australia and the UK have occurred since the 1860s and throughout the 20th century[32]. We were not able to quantitatively link the global spread of *E. coli* O157:H7 to the global cattle trade in the 19th and early 20th century as these historic data are incomplete but the importance of massive trans-atlantic cattle movements in the emergence and spread of O157 has been postulated previously[33].

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

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

Knowledge of the global spread of *E. coli* O157:H7 enables further insights into how to mitigate the effects of this pathogen and reduce the risk of future STEC emergence [34]. This study informs on how non-O157 STEC may spread globally in similar fashion as *E. coli* O157. For biosecurity and trade purposes, movements of live animals are now recorded between a number of countries (http://comtrade.un.org) and it may be sensible to test animals for STEC prior to transport, as well as animal feed for any STEC prior to transportation/shipment. This would be particularly important when a new virulent strain of STEC has been detected in a country to prevent its further spread. General test methods for detection and sequencing of *E. coli* O157:H7 and other STEC are available[35, 36]. and international collaboration will be critical here. Furthermore, computer simulations of disease emergence and transmission would help identify countries at higher risk, and inform where surveillance and control strategies for animal movements should take place[7]. Additionally, promotion of trade in germplasm would also lessen the chance of transmission.

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

Acknowledgements

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

Funding

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

Author contributions

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

EF, NS and OR wrote the manuscript.

EF, OR, BL, MM, JLB, CL, VG, RS, AH, IF, NF, TG, PB, PJ, MR, IC, JC, CJ, KG, GM, PSC, FPR, KF and NS generated and interpreted the data used in the analysis

OR, NS, CL and FPR performed the analysis

All authors reviewed the manuscript

Conflict of interest

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

Disclaimer

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

LEGENDS

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

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

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

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

Figure 5. Geographic dynamics of the time to 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 SpreaD3 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 see Fig. 7).

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

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

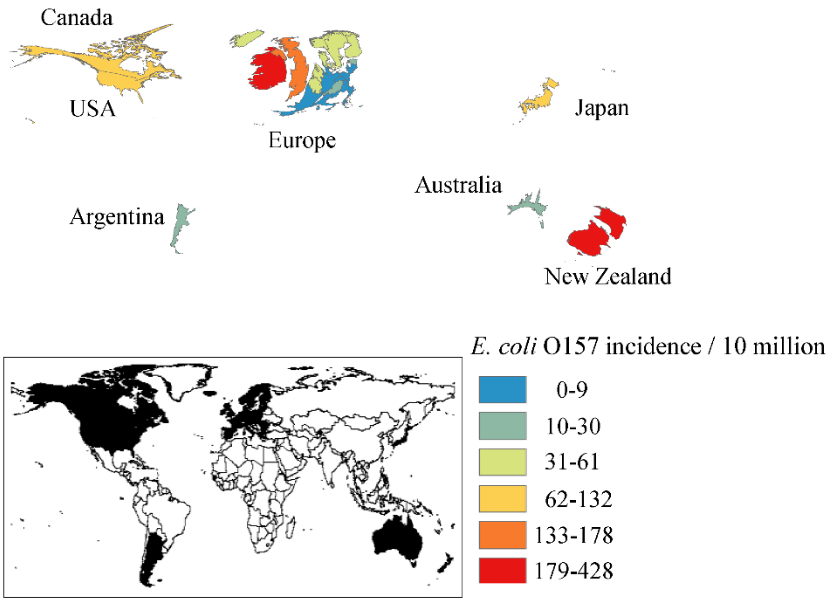


Figure 1. Cartogram of the incidence per 10 million (per year) of *E. coli* O157:H7 from 2010 – 2015 where the area of the country corresponds to the incidence (Note Australia incidence is based on the years 2001-2009 which were the only available data), and inset is a map of the world of original scale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

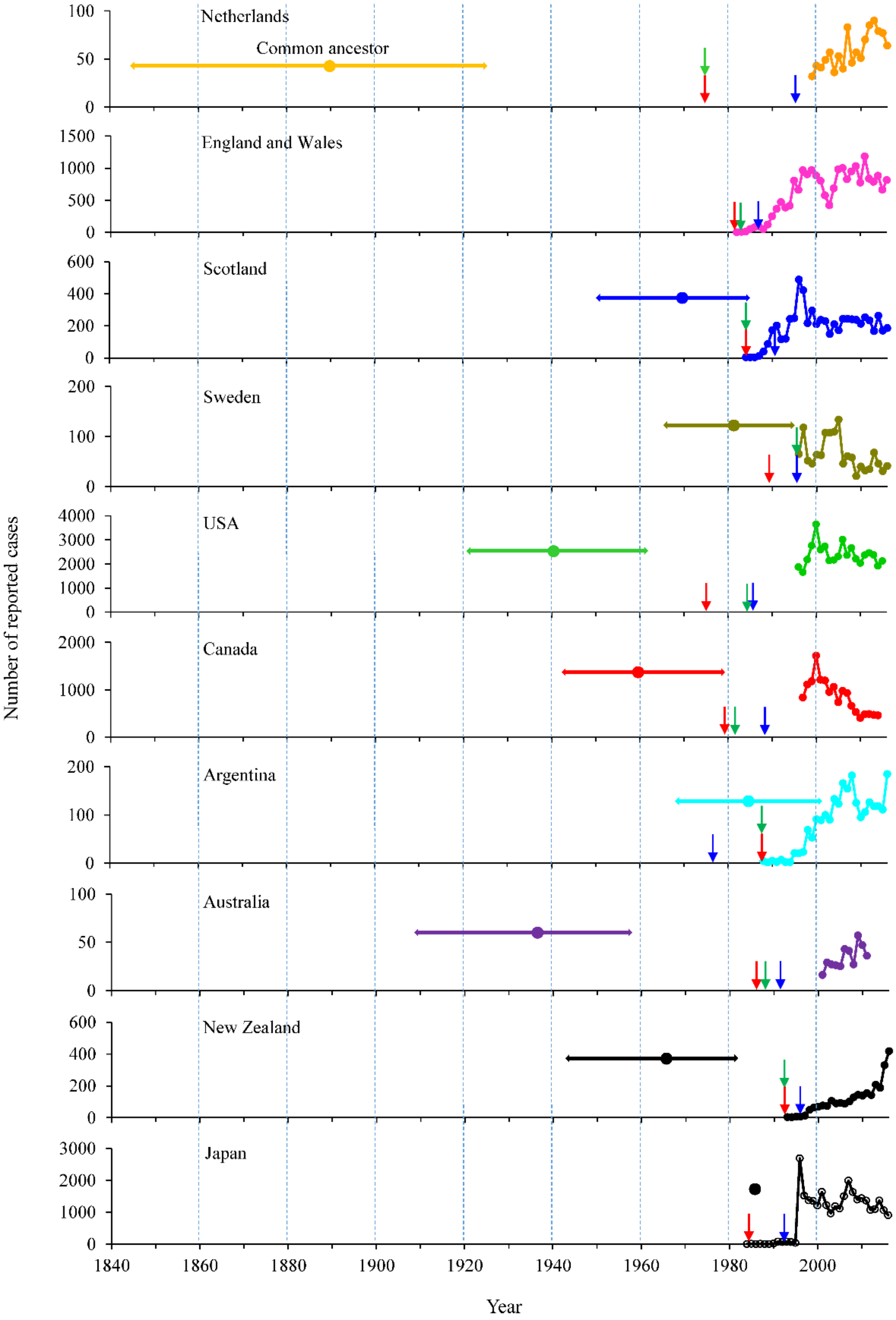


Figure 2. The reported human cases of *E. coli* O157:H7 by country obtained from national reference laboratories (red arrow indicates first human case, green arrow first human case of HUS associated with *E. coli* O157:H7 and blue arrow first isolation from cattle). The horizontal coloured bars and filled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

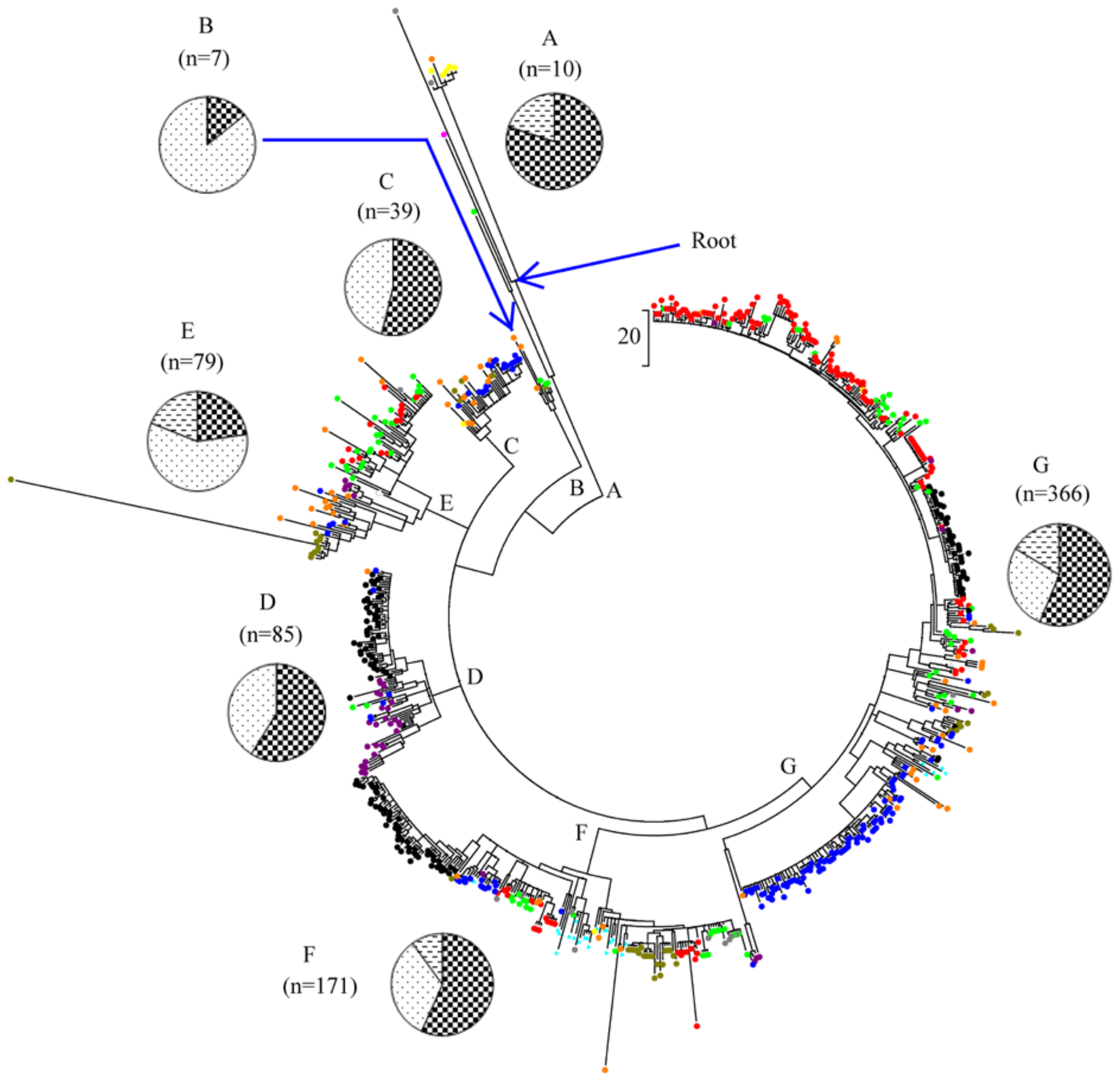
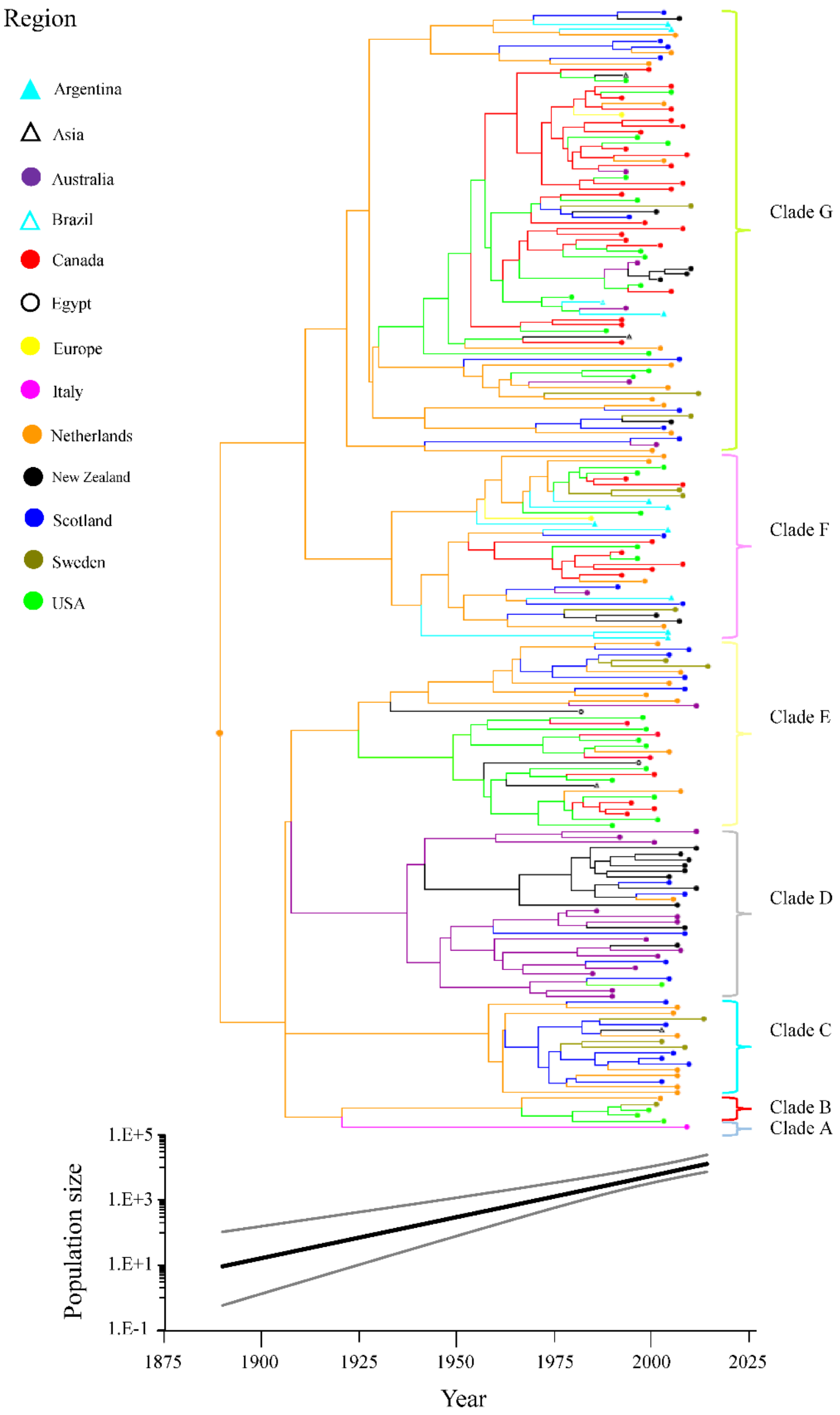


Figure 3. Nearest-neighbour joining tree of 757 *E. coli* O157:H7 isolates inferred from the 3956 SNP's obtained from PanSeq: Scotland (●); Canada (●); USA (●); the Netherlands (●); Sweden (●); New Zealand (●); Australia (●); Europe (●); Italy (●); Egypt (○); Asia (△); Argentina (▲); South America (△);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



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

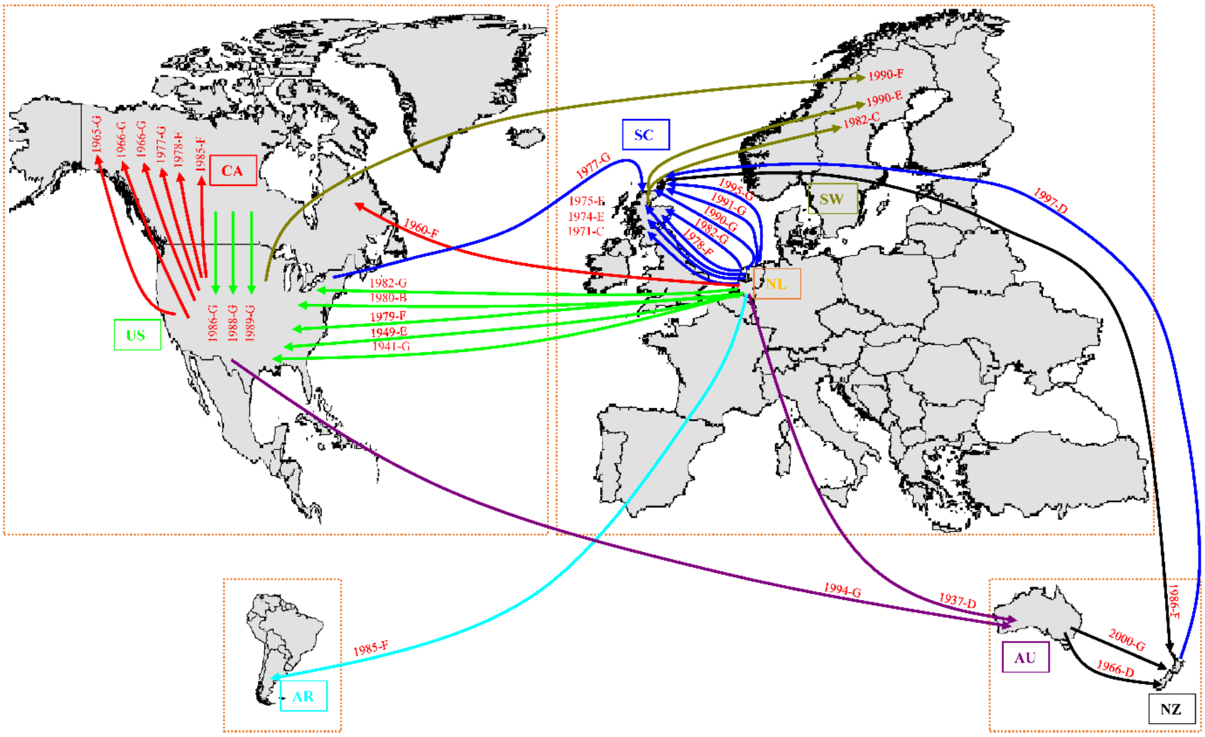


Figure 5. Geographic dynamics of the transmission of *E. coli* O157:H7. The arrows indicate 34 major transmission events and dates are the median values of the MRCA taken from BEAST. The letters denote the phylogenetic clades (see Fig. 3). The map was based on the output from SpreaD3 and reconstructed 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).

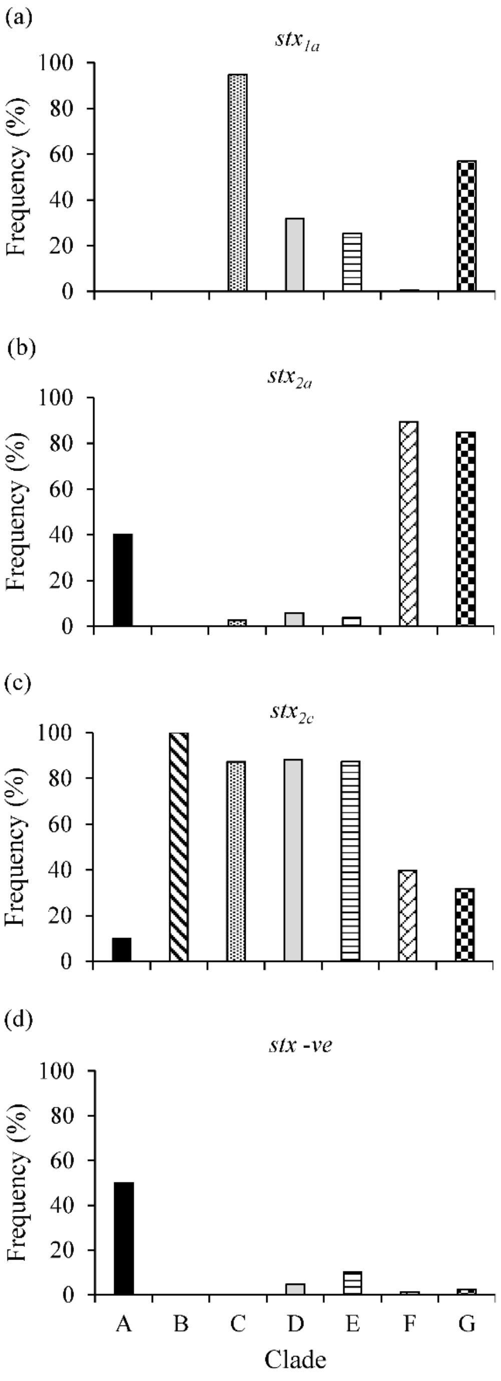


Figure 6. Frequency of *E. coli* O157:H7 Shiga toxin genes by clade for (a) *stx*$_{1a}$, (b) *stx*$_{2a}$, (c) *stx*$_{2c}$ and (d) *stx* negative.

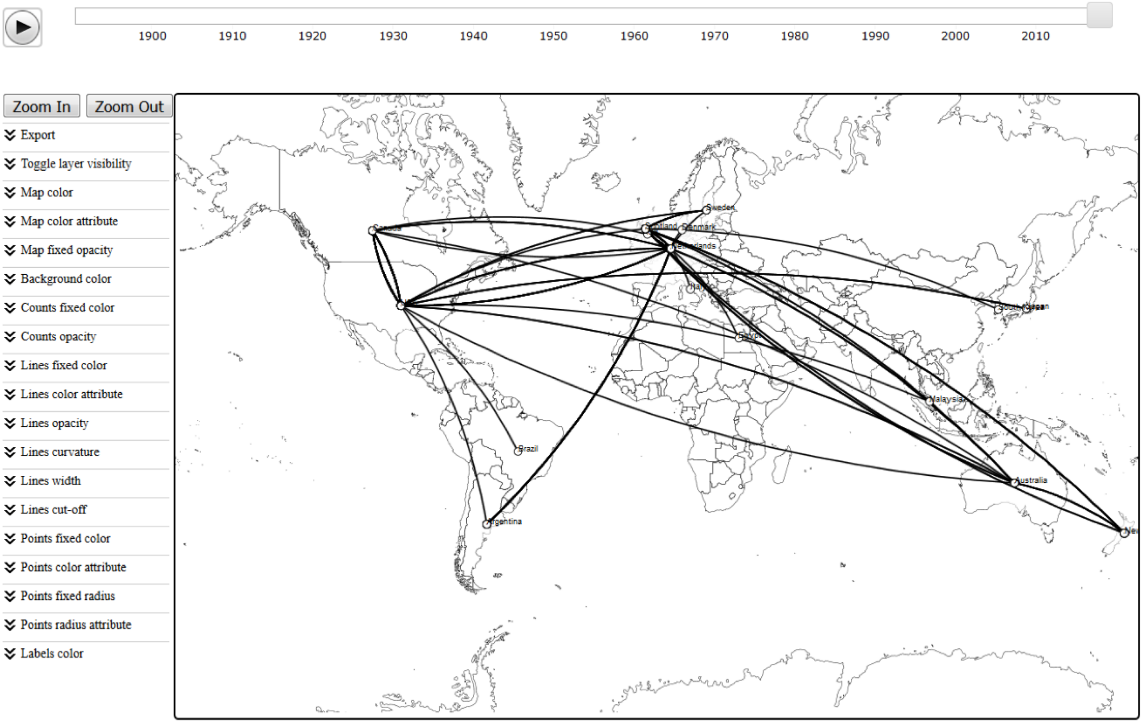


Figure 7. Screenshot of video illustrating the global spread of *E. coli* O157:H7 (Supplementary_File_S2.mp4).

References

1. Morens DM, Fauci AS. Emerging infectious diseases in 2012: 20 years after the institute of medicine report. Mbio, **2012**; 3: e00494-12.

2. Jones KE, Patel NG, Levy MA, et al. Global trends in emerging infectious diseases. Nature, **2008**; 451: 990-U4.

3. Weill F, Domman D, Njamkepo E, et al. Genomic history of the seventh pandemic of cholera in africa. Science, **2017**; 358: 785.

4. Holmes AH, Moore LSP, Sundsfjord A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. Lancet, **2016**; 387: 176-87.

5. Jones BA, Grace D, Kock R, et al. Zoonosis emergence linked to agricultural intensification and environmental change. Proc Natl Acad Sci U S A, **2013**; 110: 8399-404.

6. Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP, Daszak P. Predicting the global spread of H5N1 avian influenza. Proc Natl Acad Sci U S A, **2006**; 103: 19368-73.

7. Nelson MI, Schaefer R, Gava D, Cantao ME, Ciacci-Zanella JR. Influenza A viruses of human origin in swine, Brazil. Emerg Infect Dis, **2015**; 21: 1339-47.

8. Karmali MA, Gannon V, Sargeant JM. Verocytotoxin-producing *Escherichia coli* (VTEC). Vet Microbiol, **2010**; 140: 360-70.

9. Mughini Gras L, van Pelt W, van der Voort M, Heck M, Friesema I, Franz E. Attribution of human infections with shiga toxin- producing *Escherichia coli* (STEC) to livestock sources and identification of source- specific risk factors, the Netherlands (2010–2014). Zoonoses and Public Health, **2017**.

10. Riley L, Remis R, Helgerson S, et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. N Engl J Med, **1983**; 308: 681-5.

11. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. Super-shedding and the link between human infection and livestock carriage of *Escherichia coli* O157. Nat Rev Microbiol, **2008**; 6: 904-12.

12. Davis M, Hancock D, Besser T, et al. Correlation between geographic distance and genetic similarity in an international collection of bovine faecal *Escherichia coli* O157 : H7 isolates. Epidemiol Infect, **2003**; 131: 923-30.

13. Feng P, Lampel K, Karch H, Whittam T. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157 : H7. J Infect Dis, **1998**; 177: 1750-3.

14. Dallman TJ, Ashton PM, Byrne L, et al. Applying phylogenomics to understand the emergence of shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. Microb Genom, **2015**; 1: e000029.

15. Laing C, Buchanan C, Taboada EN, et al. Pan-genome sequence analysis using panseq: An online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics, **2010**; 11: 461,2105-11-461.

16. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the ebola epidemic. Nature, **2017**.

17. Gastner M, Newman M. Diffusion-based method for producing density-equalizing maps. Proc Natl Acad Sci U S A, **2004**; 101: 7499-504.

18. Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform, **2008**; 9: 299-306.

19. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol, **2012**; 29: 1969-73.

20. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol, **2016**; 33: 2167-9.

21. Strachan NJC, Rotariu O, Lopes B, et al. Whole genome sequencing demonstrates that geographic variation of *Escherichia coli* O157 genotypes dominates host association. Scientific Reports, **2015**; 5: 14145.

22. Wells J, Davis B, Wachsmuth I, et al. Laboratory investigation of hemorrhagic colitis outbreaks associated with a rare *Escherichia coli* serotype. J Clin Microbiol, **1983**; 18: 512-20.

23. Orskov F, Orskov I, Villar J. Cattle as reservoir of verotoxin-producing *Escherichia coli* O157:H7. Lancet, **1987**; 2: 276-.

24. Drummond AJ, Bouckaert RR. Bayesian evolutionary analysis with BEAST. 1st ed. Cambridge, United Kingdom: Cambridge University Press, **2015**.

25. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. Shiga toxin subtypes display dramatic differences in potency. Infect Immun, **2011**; 79: 1329-37.

26. Dallman TJ, Byrne L, Ashton PM, et al. Whole-genome sequencing for national surveillance of shiga toxin-producing *Escherichia coli* O157. Clin Infect Dis, **2015**; 61: 305-12.

27. Chart H, Rowe B, vd Kar N, Monnens LA. Serological identification of *Escherichia coli* O157 as cause of haemolytic uraemic syndrome in netherlands. Lancet, **1991**; 337: 437.

28. Snedeker KG, Shaw DJ, Locking ME, Prescott RJ. Primary and secondary cases in *Escherichia coli* O157 outbreaks: A statistical analysis. BMC Infect Dis, **2009**; 9: 144.

29. Dodd CC, Sanderson MW, Sargeant JM, et al. Prevalence of *Escherichia coli* O157 in cattle feeds in midwestern feedlots. Appl Environ Microbiol, **2003**; 69: 5243-7.

30. Ferens WA, Hovde CJ. *Escherichia coli* O157:H7: Animal reservoir and sources of human infection. Foodborne Pathog Dis, **2011**; 8: 465-87.

31. Houghton FL. Holstein-friesian cattle: A history of the breed and its development in america. New York, USA: Cornell University Library, **1897**.

32. Binney BM, Biggs PJ, Carter PE, Holland BM, French NP. Quantification of historical livestock importation into New Zealand 1860-1979. N Z Vet J, **2014**; 62: 309-14.

33. Leopold SR, Magrini V, Holt NJ, et al. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad Sci U S A, **2009**; 106: 8713-8.

34. Tozzoli R, Grande L, Michelacci V, et al. Shiga toxin-converting phages and the emergence of new pathogenic *Escherichia coli*: A world in motion. Front Cell Infect Microbiol, **2014**; 4: 80.

35. Franz E, Delaquis P, Morabito S, et al. Exploiting the explosion of information associated with whole genome sequencing to tackle shiga toxin-producing *Escherichia coli* (STEC) in global food production systems (vol 187, pg 57, 2014). Int J Food Microbiol, **2015**; 193: 159-.

36. Nadon C, Van Walle I, Gerner-Smidt P, et al. PulseNet international: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveill, **2017**; 22: 10.2807/1560,7917.ES.2017.22.23.30544.