Whole genome molecular epidemiology of *E. coli* O157 isolates from humans, food and the environment.

Food Standards Agency- Scotland

Contract FS102029

Final Report February 2014

Norval Strachan and Ken Forbes

UNIVERSITY OF ABERDEEN

This report has been produced by the University of Aberdeen under a contract placed by the Food Standards Agency (the Agency). The views expressed herein are not necessarily those of the Agency. The University of Aberdeen warrants that all reasonable skill and care has been used in preparing this report. Notwithstanding this warranty, the University of Aberdeen shall not be under any liability for loss of profit, business, revenues or any special indirect or consequential damage of any nature whatsoever or loss of anticipated saving or for any increased costs sustained by the client or his or her servants or agents arising in any way whether directly or indirectly as a result of reliance on this report or of any error or defect in this report.

Summary

Scotland has consistently one of the highest rates of *E. coli* O157 infection in the world. Human infection is acquired by foodborne, environmental (e.g. contact with farm animals), waterborne and person to person transmission pathways. Typing has traditionally been carried out by phage typing and pulsed field gel electrophoresis. The developments in next generation sequencing (NGS) have now made it possible to readily sequence isolates at a reasonable cost. This pilot study investigated the potential of whole genome sequencing to advance our knowledge base on the types of *E. coli* O157 strains circulating in the environment in Scotland and how these relate to the strains which are transmitted through the food chain and those which lead to illness in humans. In total 148 isolates were whole genome sequenced using an Illumina HiSeq sequencer. Shigatoxin typing was performed by in-silico PCR and phylogenetic analysis was based on core genome single nucleotide polymorphisms.

The project posed a number of key questions and the progress that was achieved in addressing these is outlined below.

- Are Scottish strains generally clustered or dispersed in the global population (e.g. comparison with USA) and is there any structure to sub-clustering of related Scottish strains?

Scottish strains are highly clustered into three broad groupings which are interspersed between strains from other regions of the world. Sub-clustering of Scottish strains was also evident.

- How similar are the strains identified in humans to those isolated from cattle and sheep? What proportion of clinical strains in Scotland can be attributed to ruminant sources?

The WGS phylogenetic tree indicates that isolates from cattle and sheep intersperse those from humans. This suggests that all isolates are from the same population and there are not any human clades where there is no ruminant isolate. The shigatoxin profiles suggest that cattle and clinical are most similar whilst there is an over-preponderance of stx2a in food/environmental whilst the opposite is the case for sheep.

- How do clinical strains with an epidemiological food link compare with the strains isolated from foodstuffs? Do these strains have identifiable characteristics which could explain their ability to get through the food chain?

The food strains are not randomly clustered throughout the phylogenetic tree. However, it should be noted that a number of the food associated strains are older than the majority of human, and ruminant strains and as such there may be confounding. There were only a handful of clinical strains that were obtained that had a food link and so it was not possible to answer this part of the question.

- What different toxin variants are circulating in animal populations in Scotland and which of these are associated with more severe disease in humans? Are there differences in strains which cause severe disease with those which do not?

More than 100 isolates carried only one shigatoxin (stx2a (76) and stx2c (27)). Twenty-six of the isolates carried a combination of stx1a with stx2a (4) or stx2c (22). From the literature Stx2a is associated with more severe disease in humans and was found in 62% human, 42% cattle, 29% sheep and 76% food/environment strains. The majority of the stx2a strains are clustered in the phylogeny indicating that these isolates have a different evolutionary history compared with other toxin types (e.g. stx1a/2c).

The significant findings of this study are: the high prevalence of the potent stx2a shigatoxin may be part of the reason for the high incidence of human disease in Scotland; the overlapping phylogeny of the cattle and sheep isolates indicates that both of these reservoirs

are likely to be important for maintenance of this organism in the farm environment; there is correlation between the shigatoxin and phage types; WGS can readily be conducted on *E. coli* O157 isolates; a robust phylogeny can be easily generated; queue times for commercial sequencing of DNA are very long (3-4 months) and software that interrogates specific aspects of the *E. coli* O157 genome is currently only partially developed.

1. Background

1.1 Introduction

Escherichia coli O157 is a gastrointestinal zoonotic pathogen of public health importance in a number of countries including Canada, the USA and Scotland [1-3]. The disease is relatively rare with the highest incidence worldwide being reported in Scotland (e.g. 4.5 cases per 100,000 in 2012[4]. *E. coli* O157 is currently characterised by phage typing, pulsed-field gel electrophoresis and detection of virulence determinants (e.g. toxins, attachment). These methods are used both in outbreak detection and monitoring of sporadic cases. However, these existing typing schemes are not sufficiently powerful or comparable to facilitate detailed epidemiological analyses to be undertaken.

The recent developments in next generation sequencing now make it affordable to whole genome sequence (WGS) bacterial genomes. It is likely that with developments in the next few years that costs will reduce further and as such it is expected that WGS will become routine for bacterial food-borne pathogen diagnosis and epidemiology. The availability of whole genome information enables a number of methods to be applied that can be used to characterise the particular isolate and to determine the genetic relatedness of isolates. These include single nucleotide polymorphism (SNP) analysis, pathotyping and multi-locus variable number tandem repeats (MLVA) [5-7].

The WGS data generated in this pilot study would represent a significant development in our knowledge base on the types of *E. coli* O157 strains circulating in the environment in Scotland and how these relate to the strains which are transmitted through the food chain and those which lead to illness in humans. This project is a pilot which will genome sequence a selection of isolates primarily from the Grampian Region of Scotland. This will determine the potential for future analyses, including source attribution, outbreak detection and pathotyping.

1.2 Aims/research Questions

The questions to be addressed through this pilot study include:

- Are Scottish strains generally clustered or dispersed in the global population (e.g. comparison with USA) and is there any structure to sub-clustering of related Scottish strains?
- How similar are the strains identified in humans to those isolated from cattle and sheep? What proportion of clinical strains in Scotland can be attributed to ruminant sources?
- How do clinical strains with an epidemiological food link compare with the strains isolated from foodstuffs? Do these strains have identifiable characteristics which could explain their ability to get through the food chain?

- What different toxin variants are circulating in animal populations in Scotland and which of these are associated with more severe disease in humans?
- Are there differences in strains which cause severe disease with those which do not?

2. Materials and Methods

2.1 Isolates

Isolates for sequencing were obtained from the collection held at the University of Aberdeen, which dates back over 20 years, as well as 12 obtained from SERL (eight food isolates and two from clinical cases associated with particular foods - cheese and sausage roll). In total 175 isolates were submitted for sequencing. Of these 148 (85%) were successfully sequenced and assembled. Table 1 provides details of the isolates sequenced.

Table 1. Isolates successfully sequenced and assembled.

Host	No successfully sequenced and assembled
Clinical	74
Cattle	26
Sheep	27
Food/Environment	21
Total	148

Table 2. Published E. coli O157 genomes.

E. coli O157 published genomes	Sequence accession number	Source	Year, Location
O157:H7 EDL933	NC_002655	Isolated from Michigan ground beef linked to an outbreak involving contaminated hamburgers.	1982, USA
O157:H7 Sakai	NC_002695	Isolated from outbreak in primary schools	1996, Japan
O157:H7 EC4115	NC_011353		
157:H7 TW14359	NC_013008	Clinical isolate obtained from a patient with EHEC infection from spinach-associated outbreak	2006, USA

2.2 Whole Genome Sequencing and Analysis

PREPARATION OF DNA FOR SEQUENCING: Growth and extraction of the *E. coli* O157 isolates was done within the CL3 facility of the Institute of Medical Sciences at the University of Aberdeen. Overnight cultures of *E. coli* O157 were grown on Harlequin SMAC-BCIG agar plates (Hal 6, Lab M, Topley House, Lancashire) at 37^{o}C. A single, well isolated sorbitol negative colony was selected and tested with *E. coli* O157 latex (code DR0620M, Oxoid, Basingstoke) and the latex plated onto Columbia Agar at 37^{o}C for 24 hours (Code CM0031, Oxoid, Basingstoke). DNA was extracted with the Wizard Genomic DNA Purification Kit (Promega UK Ltd, Southampton) as per instructions with an additional Proteinase K step (25µl, lyophilized Proteinase K reconstituted at a concentration of 6mg/260 µl Proteinase buffer) post treatment with nucleic acid lysis solution.

Concentration of DNA was determined by Picogreen assay. The DNA was then submitted to the facility at Oxford (Wellcome Trust Centre for Human Genetics (WTCHG), Oxford Genomics Centre) for WGS.

WHOLE GENOME SEQUENCING: This was conducted by WTCHG using Illumina HiSeq sequencer with 100 nt paired-end sequencing. Paired read files were supplied by WTCHG.

ASSEMBLY: The raw paired-end reads were assembled using Spades. Typically coverage was 30x and total consensus (assembled genome size) around 5.5 Mb.

Stx TOXIN TYPING: The scheme from [8] was used to determine the presence of shigatoxin subtypes. This was achieved through modification of a Perl script supplied by Chad Laing (Public Health Agency of Canada) which enabled in silico PCR's to be conducted.

PAN-GENOMIC SNP ANALYSES: Four *E. coli* O157 published genomic sequences were used (Table 2), along with all the strains sequenced in the current study to construct a pan-genome. PanSeq [9] was used to determine the non-redundant pan-genome among all strains, to determine the presence / absence of all loci and to construct multiple sequence alignments of the pan-genome. The pan-genome was constructed by using a seed genome and identifying regions of ≥1000 bp not found in the seed but present in any other genome at a 99 percent sequence identity cutoff. The pan-genome was subsequently fragmented into 1000 bp segments, and the presence / absence of each locus in every genome determined at a 99 percent sequence identity threshold. Loci present in all genomes underwent multiple sequence alignment using Muscle [10], and were concatenated together. This aligned pan-genome was used to identify SNPs in the core genome of all isolates. A phylogeny of *E. coli* O157 isolates' genomes was rooted using *E. coli* O55:H7 strain CB9615 as a proximal outgroup (O55 is considered to be the immediate ancestor of *E. coli* O157), and *E. coli* O111:H- strain 11128 as a more distal outgroup. Thus, these two strains were also included in the above core genome SNP screen.

A Neighbour Joining tree was generated in BioNumerics.

3. Results and Discussion

3.1 Investigating the relationship between shigatoxin type, source and phage type

Table 3 presents the results of the toxin typing using the scheme of Scheutz [8]. Most of the isolates are either stx2a (51%), stx2c (16%) or stx1a and stx2c (15%) positive. It was also found that 12% of the isolates did not carry any of the shigatoxins.

Table 3 Scheutz [8] shigatoxin profiles by source.

Source	stx1a &stx2a	stx1a & stx2c	stx2a	stx2c	stx2a & stx2c	Negative
Clinical	3	11	43	11	0	6
Cattle	0	6	11	6	0	3
Sheep	0	4	7	9	1	6
Food/Env	1	1	15	1	0	3

To determine whether the presence of stx2a only isolates were more or less frequent in the different sources, odds ratios were calculated and statistical significance determined by Fisher's exact test. It was found that the isolates originating from food and environmental samples were more likely to be stx2a (OR =5.87, P = 0.009) whereas the opposite was the case for sheep isolates (OR= =0.29, P = 0.01).

There have been several reports (e.g.[11]) that indicate that shigatoxin type 2 exhibit greater pathogenicity (i.e. diarrhoea plus HUS) than shiga toxin 1 isolates. It has also been recently reported [12] that stx2a are 25 times more potent than stx2c in both Vero cell and human renal proximal tubule epithelial cell assays. Further, stx2a is approximately 40-400 times more potent that stx2c and stx1 in mice [12]. Also, studies from the USA [13] found that clinical genotypes were associated with the stx2a gene. Hence, since stx2a is the most common toxin found in all sources in the current study this may contribute to the explanation of why disease incidence is so high in Grampian. The high proportion of stx2a in food and environmental sources may partly explain the fact that a number of these isolates are associated with human outbreaks. Alternatively, it could be that these isolates have improved survival in these matrices, however more data are required to evaluate this hypothesis.

It is unclear why six of the clinical strains appear to be shigatoxin negative. It could be that the shigatoxins have been lost during laboratory culture or perhaps more likely that sequence polymorphisms in the test sequences prevented their detection. Further work utilising alternative oligo sequences would determine which of these options was indeed the case.

The insertion sites of the shigatoxin phage have yet to be determined for the current dataset but this should be readily achieved by *in silico* analysis [14].

Phage types of 51 of the clinical isolates were made available by SERL and these were compared with the shigatoxin types (Table 4).

Table 4 Scheutz [8] shigatoxin profiles by Phage Type.

Phage Type	stx1a & stx2a	stx1a & stx2c	stx2a	stx2c	stx2a & stx2c	Negative
1--						1
2			2	1		
4			2			
8		3				2
21/28			26			
31				1		
32		1	1	3		3
34				1		
43			1	1		
54		1				1
RDNC						

It is apparent that PT21/28 comprise just over half of all of the isolates. All of these isolates are stx2a positive and it is worth noting that this PT has been associated with supershedding [15]. It is likely that the potency of the stx2a shigatoxin combined with the supershedding makes this a particularly important phage type in terms of human disease.

Only one of the PT32 isolates appeared to carry the stx2a shigatoxin by the in silico PCR. This is in agreement with other research in Scotland [16] which shows a high relative proportion of stx2c in this PT which is also associated with low shedding.

The paired shigatoxins, stx1a plus stx2c, appear to be associated with PT8 but the numbers of isolates are small so it is difficult to make definitive conclusions.

3.2 Food and Environment Isolates

Table 5 provides details of the food and environmental isolates. These included two isolates which were from human cases that were associated with particular foods (cheese and sausage roll). Ten of these isolates were supplied by SERL along with the phage typing information. The most common types were PT21/28 (n=5) and PT32(n=2) and PT2 (n=2). According to the Scheutz scheme 16 of the isolates were stx2a, 4 were stx –'ve and one each of stx2c, stx1a & 2a and stx1a & 2c.

Table 5 Food and Environment Isolates

Isolate Number	SERL	Original Aberdeen Number	Phage Type	Shigatoxins	Source
G163	Y	736	21/28	Stx2a	case related to cheese[a]
G164	Y	864	21/28	Stx2a	case related to sausage roll[a]
G165	Y	1209	21/28	Stx2a	goat's milk
G166	Y	2147	2	Stx2a	Ham joint, Central Scotland
G167	Y	3351	2	Stx2a	Beef burger
G168	Y	3362	21/28	Stx2a	Food - public health
G169	Y	3368	54	-'ve	meatball
G170	Y	3580	32	Stx2a	sausage
G171	Y	3581	32	Stx2a	burger
G172	Y	3617	21/28	Stx2a	beef burger
G95	N	907(1)-230		Stx2a	Cheese
G96	N	935-301		Stx2a	Cheese (NE outbreak)
G91	N	E771(13)		Stx2a	Cheese (Greig)
G93	N	E850(17)		Stx2a	Mains water Midlothian
G99	N	308		-'ve	Mince from supermarket
G160	N	TR2		Stx2c	Mince
G98	N	307		-'ve	Mince from butcher
G97	N	306		-'ve	Mince from butcher
G173	N	TR-1		Stx2a	Mince
G90	N	E367(7)		Stx2a	Raw milk Midlothian
G94	N	E867(19)		Stx2a	Uncooked sausage roll
G92	N	E798(15)		Stx1a & Stx2a	Water sample Aberdeen
G11	N			Stx1a & Stx2c	Unknown

[a] Isolate included because of the linkage of the human case to a food source

3.3 Phylogeny of E. coli O157 isolates

A phylogenetic tree of the *E. coli* O157 isolates was generated utilising the 8559 SNPs obtained from PanSeq (Figs. 1(a)-(e)). These figures are best viewed printed in A3 format.

Fig 1(a) shows the phylogeny with source information. It can be seen that isolates from cattle (brown), sheep (grey), clinicals (pink) and food (green) are distributed across the tree. This suggests that in general there is no strong association of source or host species to particular clades of *E. coli* O157. Since there are no cattle or sheep only clades or clades from which human cases are absent it follows that from across the whole phylogeny strains have the potential of causing infection in humans.

Fig 1(b) shows the relationship between phylogeny and shigatoxin type. Here there is a fairly large amount of segregation between the different shigatoxin types. In particular the shigatoxin stx 1a & 2c variant (green) is found towards the top of the tree, whereas the stx2a variant is found throughout the bottom two-thirds of the tree. The stx2c variant (blue) is more widely distributed but there is a cluster of nine strains about one quarter of the way down the tree. The PanSeq method utilises SNPs from the "core" genome, of all 151 isolates, to infer phylogeny. Hence, it is very unlikely that any of the shigatoxin phage contribute SNPs in the current analysis. This suggests that there must be other loci of the *E. coli* O157 genome that are associated with shigatoxin type. The SNPs which are most informative in predicting this association may be in areas of the genome that are important with this regard. However, there may also be genetic elements in the accessory genome that are important and this has yet to be investigated.

Fig 1(c) displays the relationship between phylogeny and phage type. There is some evidence of clustering with PT32 (yellow) being predominantly in one cluster, PT21/28 present in the bottom half of the tree and PT2 being at the bottom of the tree.

In Fig 1(d) the date of isolation appears not to correlate with phylogeny.

All isolates had unique genotypes, even within the strict criteria set for the selection of core genome, and could therefore be distinguished from each other. Indeed of the 8559 SNP site identified, 90% of these were isolate unique, and 871 SNPs were phylogenetically informative. The removal of these isolate specific SNPs allowed the identification of multi-isolate strains (Figure 1(e)). Further analysis of these isolate clusters will shed light on both epidemiological relationships, including outbreaks, and on the robustness of WGS SNP analysis compared to the current gold standard analyses of PFGE and MLVA.

It was not possible within the time available to conduct a comprehensive comparison with international isolates. This could be achieved to a limited extent using the 25-plus whole genome sequences that are publicly available. More comprehensively in terms of countries but with considerably less resolution this could be achieved by LSPA6 typing [17]

3.4 Are Scottish strains generally clustered or dispersed in the global population?

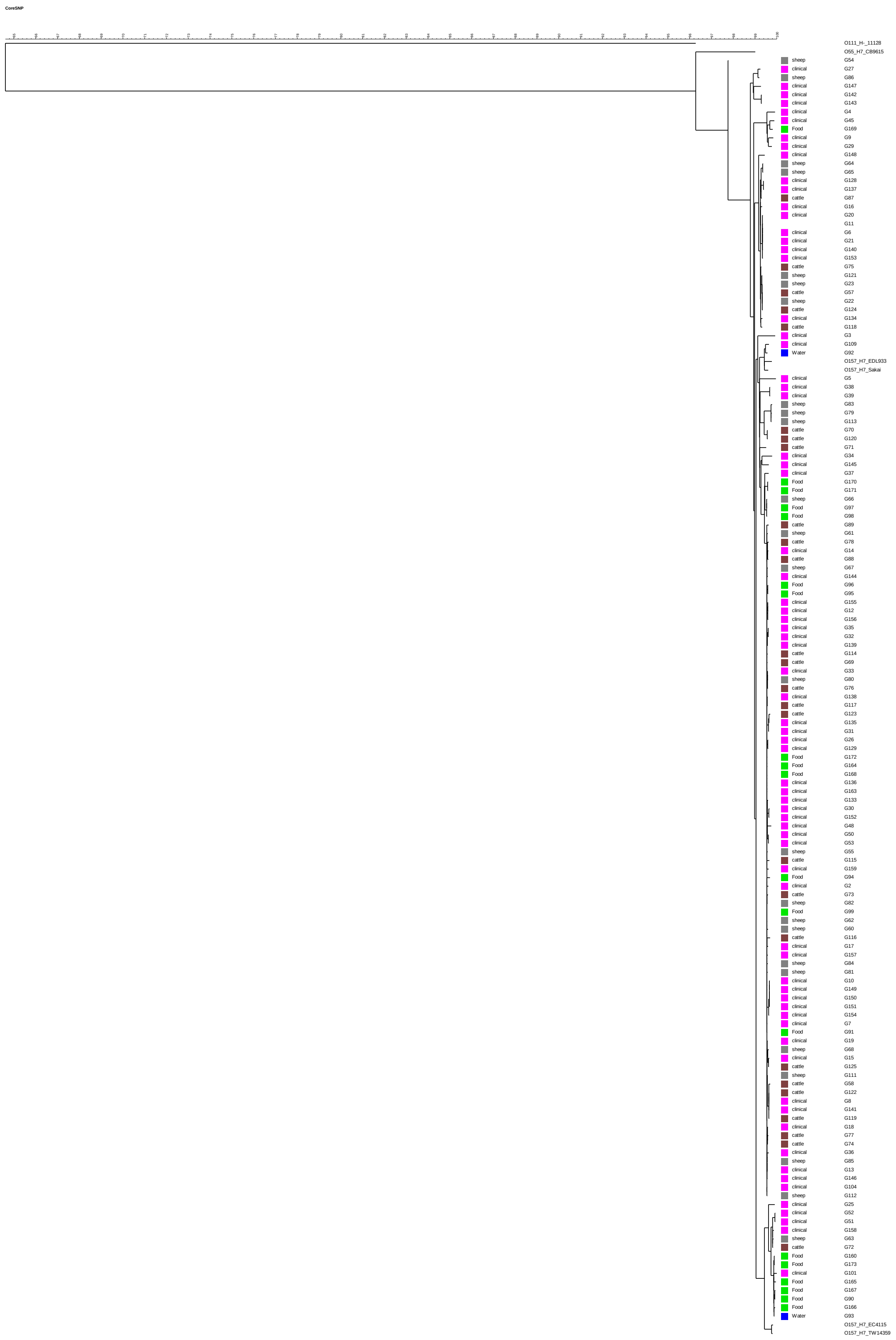
The relationship of Scottish isolates of O157 to those isolated from elsewhere in the world was explored. WGS of 66 isolates from NCBI and from Patric.org were downloaded and included along with the Scottish isolate WGS in a dedicated PanSeq analysis using the standard settings of this study. The resultant 4728 SNPs were used to construct a Neighbour Joining tree using the standard settings of this study. The resultant tree, circularised for ease of representation, is illustrated in Figure 2.

It is immediately apparent that the Scottish strains are highly clustered into three broad groupings which are interspersed between strains from other regions of the world. These foreign isolates are dominated by those from the USA, and indeed these also seem to be showing distinct clustering, as to a lesser extent do isolates from Asia. It would be of interest to know whether the few Scottish isolates in apparently 'foreign' clusters had been acquired abroad.

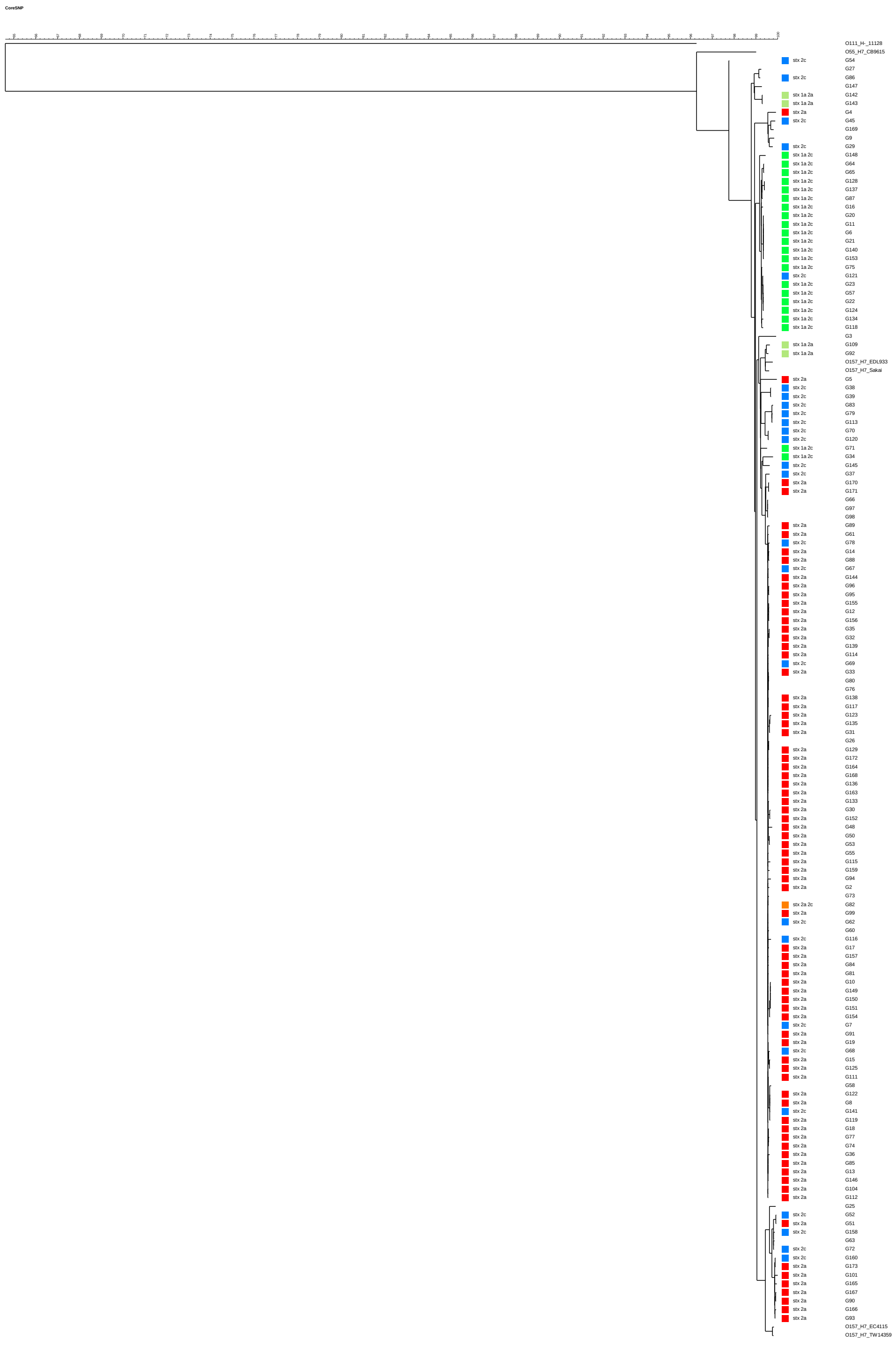
Certainly within the Scottish collection there is evidence that there has been dramatic clonal expansion of a limited number of lineages, which are further characterised by toxin and phage typing characters. This view is suggestive that there is largely a circulation of strains within Scotland, with little – but not zero – import from abroad.

Figure 1. Phylogeny of *E. coli* O157 indicating (a) source, (b) shigatoxin type, (c) phage type, (d) date, (e) multi-isolate strains.

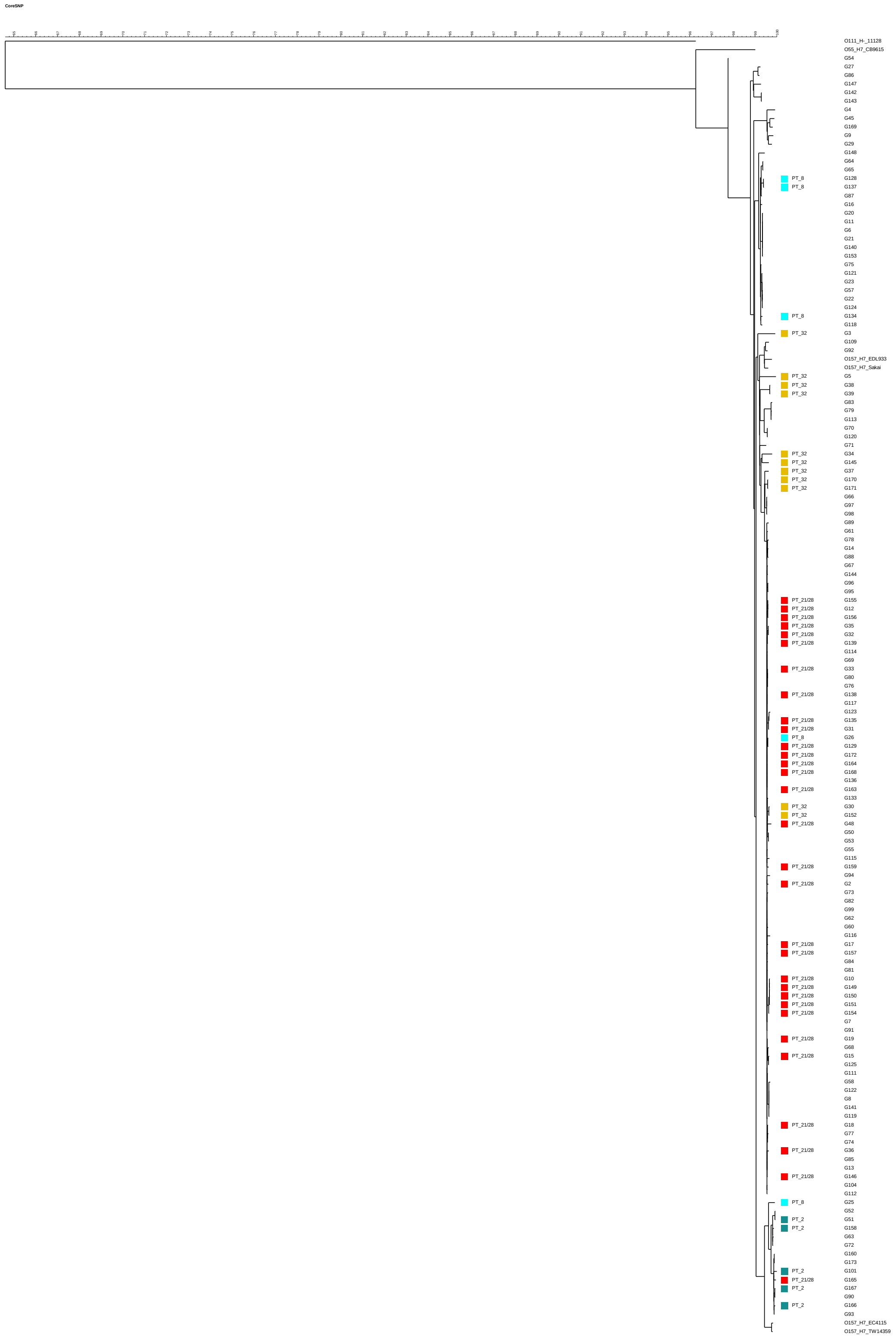
(a) Source



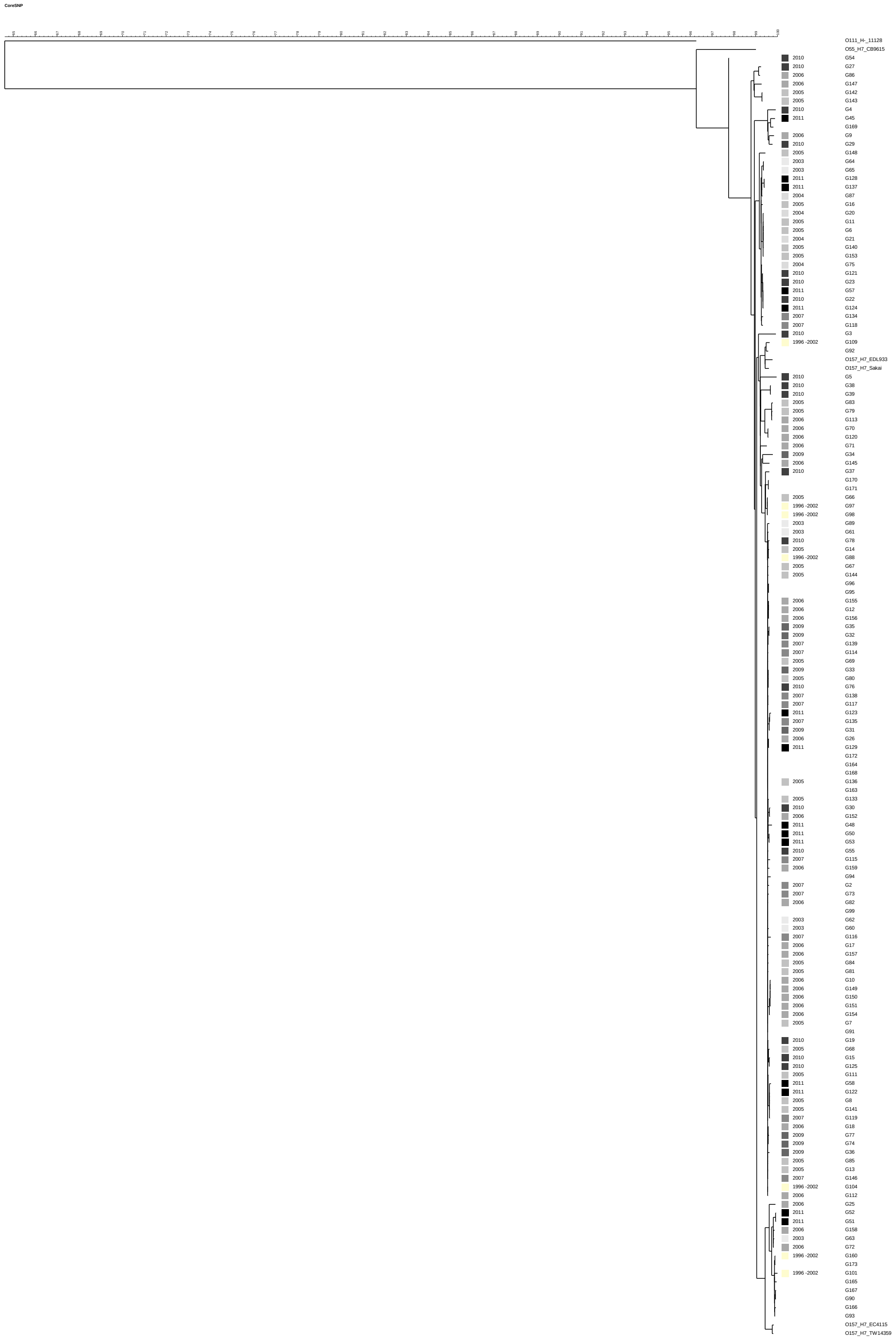
(b) Shigatoxin type



(c) Phage Type


CoreSNP
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
O111_H-_11128
O55_H7_CB9615
G54
G27
G86
G147
G142
G143
G4
G45
G169
G9
G29
G148
G64
G65
PT_8
G128
PT_8
G137
G87
G16
G20
G11
G6
G21
G140
G153
G75
G121
G23
G57
G22
G124
PT_8
G134
G118
PT_32
G3
G109
G92
O157_H7_EDL933
O157_H7_Sakai
PT_32
G5
PT_32
G38
PT_32
G39
G83
G79
G113
G70
G120
G71
PT_32
G34
PT_32
G145
PT_32
G37
PT_32
G170
PT_32
G171
G66
G97
G98
G89
G61
G78
G14
G88
G67
G144
G96
G95
PT_21/28
G155
PT_21/28
G12
PT_21/28
G156
PT_21/28
G35
PT_21/28
G32
PT_21/28
G139
G114
G69
PT_21/28
G33
G80
G76
PT_21/28
G138
G117
G123
PT_21/28
G135
PT_21/28
G31
PT_8
G26
PT_21/28
G129
PT_21/28
G172
PT_21/28
G164
PT_21/28
G168
G136
PT_21/28
G163
G133
PT_32
G30
PT_32
G152
PT_21/28
G48
G50
G53
G55
G115
PT_21/28
G159
G94
PT_21/28
G2
G73
G82
G99
G62
G60
G116
PT_21/28
G17
PT_21/28
G157
G84
G81
PT_21/28
G10
PT_21/28
G149
PT_21/28
G150
PT_21/28
G151
PT_21/28
G154
G7
G91
PT_21/28
G19
G68
PT_21/28
G15
G125
G111
G58
G122
G8
G141
G119
PT_21/28
G18
G77
G74
PT_21/28
G36
G85
G13
PT_21/28
G146
G104
G112
PT_8
G25
G52
PT_2
G51
PT_2
G158
G63
G72
G160
G173
PT_2
G101
PT_21/28
G165
PT_2
G167
G90
PT_2
G166
G93
O157_H7_EC4115
O157_H7_TW14359

(d) Date of Isolation



(e) Multi-isolate strains

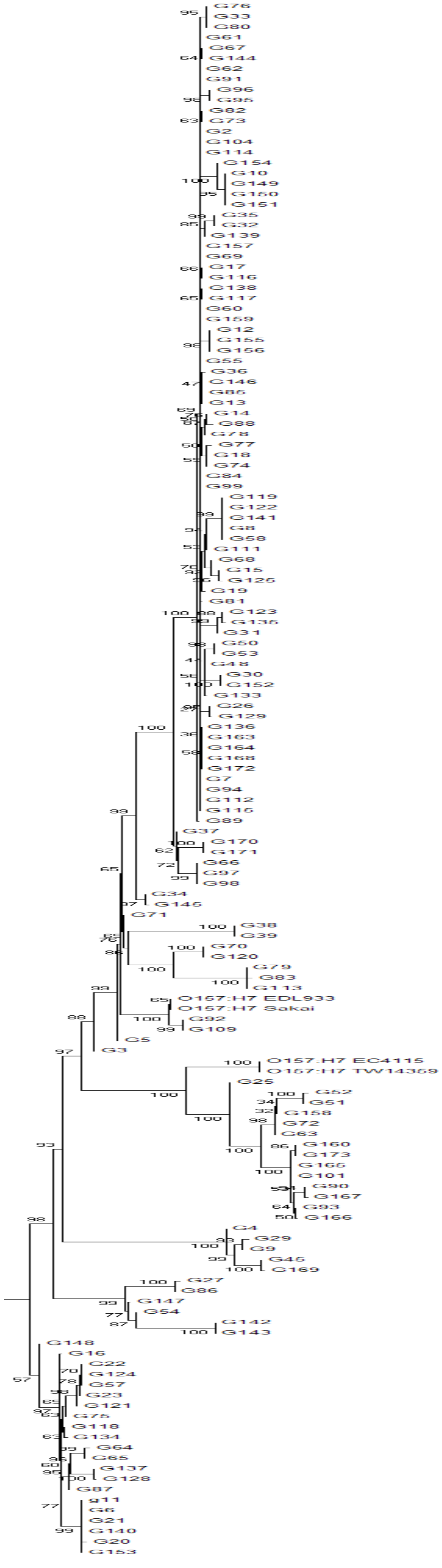


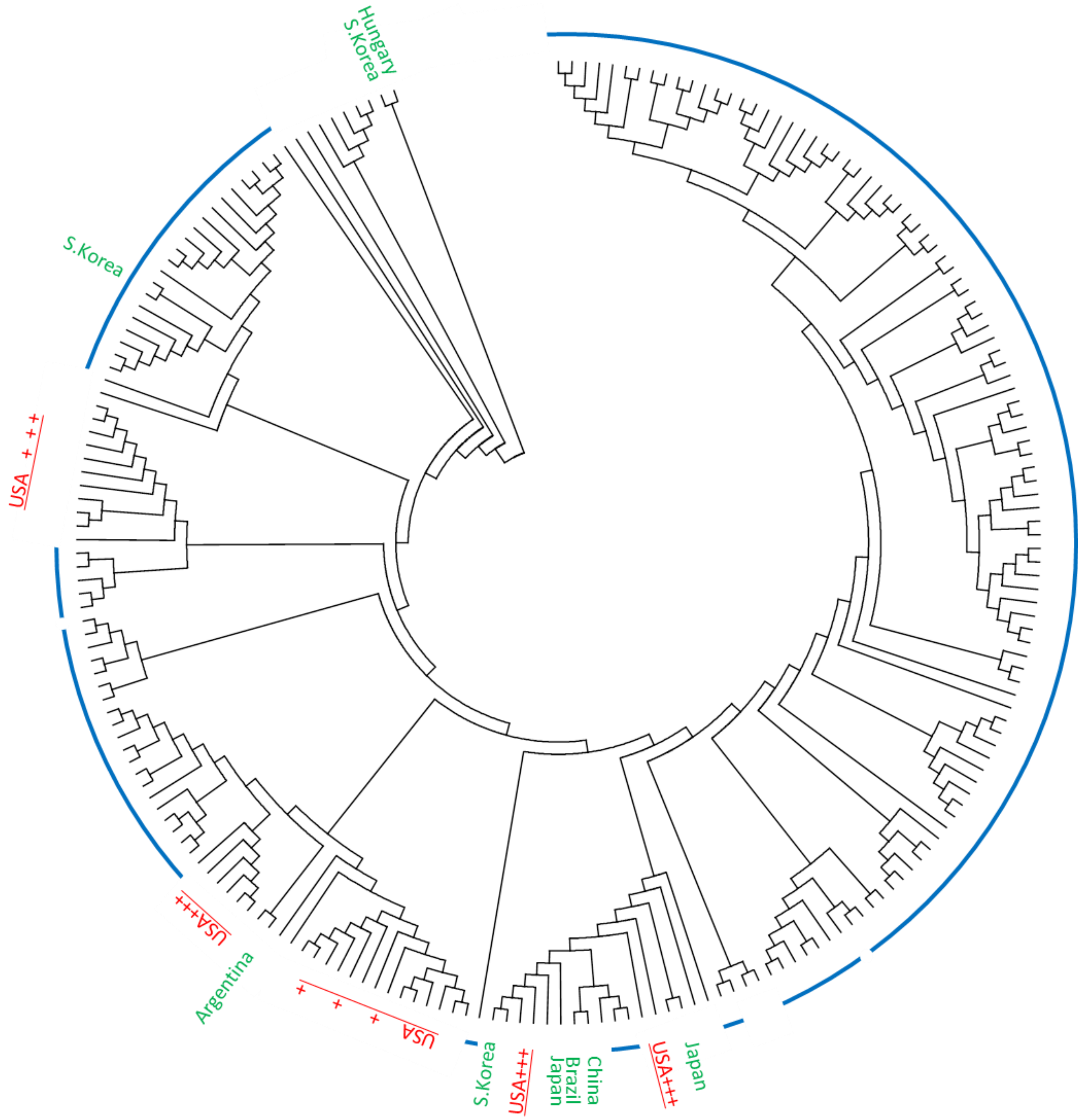
Figure 2. Phylogeny of *E. coli* O157 indicating Scottish and non-Scottish isolates.

Isolates from:

Scotland

USA (with indication of abundance of other isolates not included)

Other Countries



3.5 Identification of two non-E. coli O157 isolates

Two of the isolates, originating from sheep (G161 and G162) had been previously considered to be *E. coli* O157. They were isolated on CT-SMAC, were sorbitol negative, gave a positive reaction by latex agglutination for *E. coli* O157 but were shigatoxin negative. It became clear during the PanSeq analysis that these two isolates were outliers to the main *E. coli* O157 clade. A further PanSeq analysis was performed that included these two genomes, as well as eleven other *E. coli* serotypes and *S. typhimurium* as out group. It can be readily seen (Fig. 3) that these genomes cluster with *E. fergusonii. E. Fergusonii* is considered to be an opportunistic pathogen in both humans and animals [18], and is typically sorbitol negative.

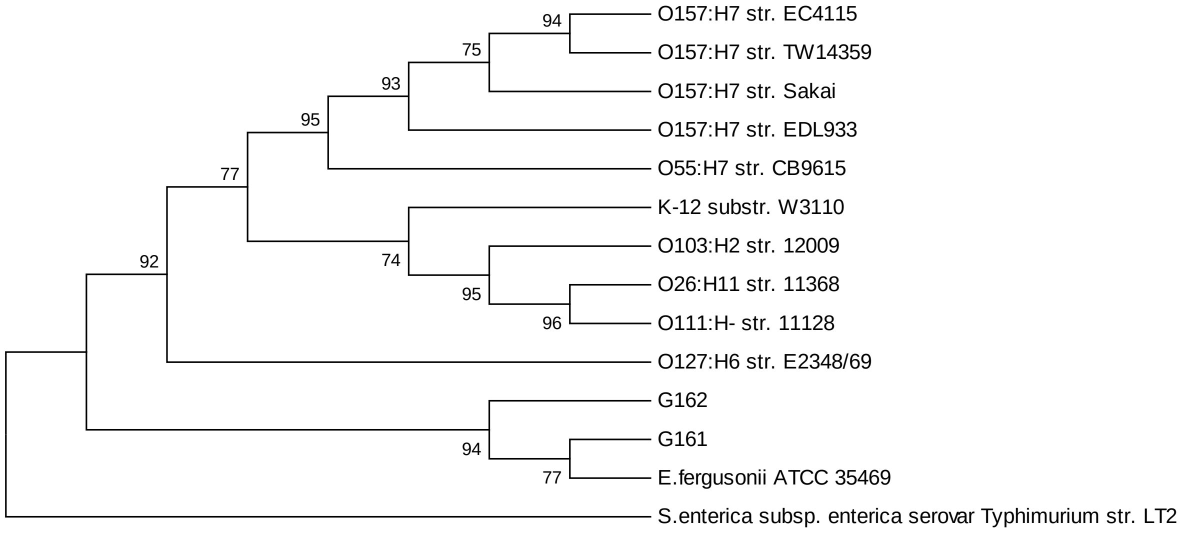


Figure 3. Phylogeny of 11 *E. coli* serotypes with two outlying putative *E. coli* O157 genomes (G161 and G162) and *S. typhimurium* as out group.

4. Conclusions

There are four main conclusions from this work:

- Isolates from clinicals, cattle and sheep appear to be distributed throughout the phylogeny of *E. coli* O157. This suggests that *E. coli* O157 is circulating between both cattle and sheep, both of which are potential reservoirs of infection in humans.
- The supershedding PT21/28 carries the most potent shigatoxin (stx2a)
- There appear to be associations between shigatoxin genes, phage types and phylogeny.
- Multi-isolate strains can be identified readily.

The analysis conducted on this dataset of WGS's should be seen as preliminary. More detailed analyses will include identification of the location of the shigatoxin phage within the genomes, comparison with genomes from other countries, detection of antimicrobial resistance genes, investigation of the accessory genome and the exploration of multi-isolate strains.

5. Acknowledgements

Dr's Lesley Alison and Mary Hanson from SERL for providing *E.* coli O157 isolates of foodborne origin. Dr Vic Gannon and Dr Chad Laing from the Public Health Agency of Canada for carrying out the assembly of the genomes using SPADES and providing a Perl script that facilitates in silico PCR for shigatoxins. Aberdeen Royal Infirmary for supply of clinical isolates.

6. References

1. Kretzschmar M, Gomes MGM, Coutinho RA, Koopman JS. (2010) Unlocking pathogen genotyping information for public health by mathematical modeling. Trends Microbiol 18: 406-412.

2. Locking ME, Pollock KGJ, Allison LJ, Rae L, Hanson MF, et al. (2011) Escherichia coli O157 infection and secondary spread, scotland, 1999-2008. Emerging Infectious Diseases 17: 524-527.

3. Pennington H. (2010) Escherichia coli O157. Lancet 376: 1428-1435.

4. Locking M, Browning L, Smith-Palmer A, Brownlie S. (2013) Gastro-intestinal and foodborne infections: Incidence of *E. coli* O157, *salmonella* and *campylobacter* reported to HPS: 2012. Health Protection Scotland Weekly Report 47: 44-45.

5. Bono JL, Smith TPL, Keen JE, Harhay GP, McDaneld TG, et al. (2012) Phylogeny of shiga toxin-producing escherichia coli O157 isolated from cattle and clinically ill humans. Mol Biol Evol 29: 2047-2062.

6. Lobersli I, Haugum K, Lindstedt B. (2012) Rapid and high resolution genotyping of all escherichia coli serotypes using 10 genomic repeat-containing loci. J Microbiol Methods 88: 134-139.

7. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW, et al. (2008) Variation in virulence among clades of escherichia coli O3157 : H7, associated with disease outbreaks. Proc Natl Acad Sci U S A 105: 4868-4873.

8. Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, et al. (2012) Multicenter evaluation of a sequence-based protocol for subtyping shiga toxins and standardizing stx nomenclature. J Clin Microbiol 50: 2951-2963.

9. Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, et al. (2010) Pan-genome sequence analysis using panseq: An online tool for the rapid analysis of core and accessory genomic regions. BMC Bioinformatics 11: 461-2105-11-461.

10. Edgar RC. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

11. Liu F, Huang J, Sadler JE. (2011) Shiga toxin (stx)1B and Stx2B induce von willebrand factor secretion from human umbilical vein endothelial cells through different signaling pathways. Blood 118: 3392-3398.

12. Fuller CA, Pellino CA, Flagler MJ, Strasser JE, Weiss AA. (2011) Shiga toxin subtypes display dramatic differences in potency. Infect Immun 79: 1329-1337.

13. Shringi S, Schmidt C, Katherine K, Brayton KA, Hancock DD, et al. (2012) Carriage of stx2a differentiates clinical and bovine-biased strains of escherichia coli O157. PLoS One 7: e51572.

14. Mellor GE, Sim EM, Barlow RS, D'Astek BA, Galli L, et al. (2012) Phylogenetically related argentinean and australian escherichia coli O157 isolates are distinguished by virulence clades and alternative shiga toxin 1 and 2 prophages. Appl Environ Microbiol 78: 4724-4731.

15. Chase-Topping M, Gally D, Low C, Matthews L, Woolhouse M. (2008) Super-shedding and the link between human infection and livestock carriage of escherichia coli O157. Nat Rev Microbiol 6: 904-912.

16. Matthews L, Reeve R, Gally DL, Low JC, Woolhouse ME, et al. (2013) Predicting the public health benefit of vaccinating cattle against escherichia coli O157. Proc Natl Acad Sci U S A 110: 16265-16270.

17. Yang Z, Kovar J, Kim J, Nietfeldt J, Smith DR, et al. (2004) Identification of common subpopulations of non-sorbitol-fermenting, beta-glucuronidase-negative escherichia coli O157:H7 from bovine production environments and human clinical samples. Appl Environ Microbiol 70: 6846-6854.

18. Wragg P, La Ragione RM, Best A, Reichel R, Anjum MF, et al. (2009) Characterisation of escherichia fergusonii isolates from farm animals using an escherichia coli virulence gene array and tissue culture adherence assays. Res Vet Sci 86: 27-35.

ANNEX A Limitations of current genotyping methods and advantages of WGS

Current genotyping methods may have:

- inadequate levels of discrimination
- limited availability of reagents
- lack of inter-laboratory standardisation
- poor reproducibility within and between laboratories
- different methods of classifying the relatedness between isolates

WGS technology:

- allows the determination of the entire, or vast majority of, genetic capacity of an isolate in one experiment
- substantially reduces, or even eliminates, a requirement for repeat testing, vastly reducing the possibility of errors

WGS allows:

- determination of laboratory cross-contamination
- authentication of relapse or reinfection in cases of a second episode of infection
- epidemiological surveillance and public-health decisions for community-acquired outbreaks, nosocomial outbreaks, and bioterrorist attacks through higher resolution typing
- investigations of bacterial population biology and evolution using phylogenies generated using information from a significant proportion of the genome
- improved correlation of particular strains (e.g. clinical isolates) to actual sources (e.g. poultry producers)
- improved source attribution by using more and different loci
- Virulence factor characterisation (e.g. presence of shigatoxins) can be performed in-silico without the need of further wet biology (i.e. PCR and sequencing)
- identify genes that are associated with human illness e.g. bloody diarrhoea, Guillain-Barré Syndrome, arthritis, etc. Identify new virulence genes, potential targets for risk assessment and intervention strategies, development of rapid screening tests based on relevant genes
- improved understanding of host–bacterial interactions at level of commensals and as pathogens.
- develop specific culture media or media selective for particular strains using knowledge of biochemical pathways.
- detection of antimicrobial resistances (but does not enable determination of whether the phenotype is expressed).
- design new antimicrobials and identification of antimicrobial targets.

Prior to WGS it would not be possible to:

- generate information on the virulence genes in one simple operation
- generate a robust phylogenetic tree that is not based on a few perhaps biased genetic characters (e.g. MLVA) or lacks informational depth (e.g. PFGE and MLVA)
- correlate between characters (stx and phage types, geography etc.) and specific lineages
- ensure future-proofing for analysis methods yet to be developed