MICROBIAL GENOMICS

RESEARCH ARTICLE

Ingle *et al.*, *Microbial Genomics* 2019;5

DOI 10.1099/mgen.0.000268

# Emergence and divergence of major lineages of Shiga-toxin-producing *Escherichia coli* in Australia

Danielle J. Ingle[1,2], Anders Gonçalves da Silva[1], Mary Valcanis[1], Susan A. Ballard[1], Torsten Seemann[1,3], Amy V. Jennison[4], Ivan Bastian[5], Rolf Wise[5], Martyn D. Kirk[2], Benjamin P. Howden[1,6] and Deborah A. Williamson[1,*]

**Abstract**

Shiga-toxin-producing *Escherichia coli* (STEC) infection is an important global cause of foodborne disease. To date however, genomics-based studies of STEC have been predominately focused upon STEC collected in the Northern Hemisphere. Here, we demonstrate the population structure of 485 STEC isolates in Australia, and show that several clonal groups (CGs) common to Australia were infrequently detected in a representative selection of contemporary STEC genomes from around the globe. Further, phylogenetic analysis demonstrated that lineage II of the global O157:H7 STEC was most prevalent in Australia, and was characterized by a frameshift mutation in *flgF*, resulting in the H-non-motile phenotype. Strong concordance between *in silico* and phenotypic serotyping was observed, along with concordance between *in silico* and conventional detection of *stx* genes. These data represent the most comprehensive STEC analysis from the Southern Hemisphere, and provide a framework for future national genomics-based surveillance of STEC in Australia.

## DATA SUMMARY

1. All sequence data have been deposited in GenBank; accession number: BioProject PRJNA319594.

## INTRODUCTION

Shiga-toxin-producing *Escherichia coli* (STEC) infection is an important cause of foodborne disease, with an estimated 2.48 million cases globally in 2010 [1]. Symptoms range from mild gastroenteritis to severe bloody diarrhoea [2], and approximately 3–6 % of cases may develop haemolytic uremic syndrome (HUS) [3, 4]. Many countries have reported STEC outbreaks as a result of ingestion of contaminated food or water, and direct or indirect contact with human or animal carriers [5–7]. Increasing globalization of the food manufacturing and supply chain means that STEC outbreaks have the potential to cross international borders, illustrated by the large multi-national 2011 European Shiga-toxigenic enteroaggregative *E. coli* O104:H4 outbreak associated with contaminated fenugreek sprout seeds [8, 9].

Over the past 8 years, whole genome sequencing (WGS) has been increasingly used for public health surveillance and outbreak investigations, particularly for foodborne pathogens [10]. Recent work from the UK has demonstrated the feasibility of WGS-based national surveillance of STEC, and highlighted the ability of WGS to detect clusters of STEC cases that were not identified by conventional epidemiological and laboratory-based investigations [5, 10, 11]. Moreover, additional work has demonstrated strong concordance between conventional STEC serotyping for the lipopolysaccharide (LPS) O-antigen and flagellar H-antigen, and *in silico*-derived serotyping, further adding to the utility of WGS-based approaches to STEC analysis in a public health microbiology setting [11–14].

Like other countries, the main reservoirs of STEC in Australia are healthy ruminants, particularly cattle [15]. In 2016, there

Received 19 August 2018; Accepted 25 March 2019; Published 20 May 2019

**Author affiliations:** [1]Microbiological Diagnostic Unit Public Health Laboratory at the University of Melbourne, The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia; [2]National Centre for Epidemiology and Population Health, The Australian National University, Canberra, Australia; [3]Melbourne Bioinformatics Group, Victoria, Australia; [4]Public Health Microbiology, Forensic and Scientific Services, Queensland Department of Health, Queensland, Australia; [5]SA Pathology, South Australia, Australia; [6]Doherty Applied Microbial Genomics, Department Microbiology and Immunology, The University of Melbourne, The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia.

***Correspondence:** Deborah A. Williamson, deborah.williamson@unimelb.edu.au

**Keywords:** epidemiology; enteric pathogens; STEC; genomic epidemiology; evolution.

**Abbreviations:** CIDT, culture independent diagnostic testing; HUS, haemolytic uremic syndrome; MDU PHL, The Microbiological Diagnostic Unit Public Health Laboratory; STEC, Shiga-toxin-producing *Escherichia coli*; WGS, whole genome sequencing.

All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary figures and three supplementary tables are available with the online version of this article.

were 341 notified cases of STEC infection in Australia [16], and a previous study suggested that STEC infections cost the Australian economy approximately $2.6 million annually [17]. Although most cases of STEC infection in Australia are considered sporadic, major STEC outbreaks can occur, with the largest reported outbreak of serogroup O157 STEC in Australia occurring amongst 57 patients in Queensland in 2013, associated with exposure to an animal nursery at an agricultural show [18]. However, at present in Australia, there is no national laboratory-based surveillance of STEC infections, resulting in a major knowledge gap in our understanding of STEC epidemiology, hampering prevention and control efforts. In Australia, epidemiological approaches for STEC surveillance have been confined to conventional serotyping, providing limited ability to differentiate sporadic infections from potential point-source outbreaks [10]. Accordingly, there is a pressing need for high-resolution characterization of STEC in Australia.

Here, we used WGS to determine the population structure of Australian STEC isolates from a spatially and temporally diverse national collection. To better inform genomic surveillance of STEC, we aimed to identify the major lineages responsible for STEC infections in Australia and to compare the utility and feasibility of *in silico* typing and virulence gene detection with conventional typing approaches. Our data represent the most comprehensive STEC dataset to date from the Southern Hemisphere and provide a valuable resource for future surveillance studies of STEC both within Australia and internationally.

**IMPACT STATEMENT**

The emergence and spread of Shiga-toxin-producing *Escherichia coli* (STEC) has been well described in Europe and North America. However, comparatively little is known about the circulation of STEC in the Southern Hemisphere. Here, in the largest STEC analysis to date from the Southern Hemisphere, we describe the major lineages of STEC in Australia, and further demonstrate the public health utility of genomic data for the characterization and surveillance of STEC. Our data provide a valuable framework for ongoing national genomics-based surveillance of this important public health pathogen.

## METHODS

### Data sources, sampling strategy and microbiological testing

In Australia, STEC is a notifiable disease, and diagnostic laboratories are requested to forward STEC isolates to a reference laboratory for further characterization. The Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL) is the public health reference laboratory for the State of Victoria in Australia and is also the national reference laboratory for epidemiological typing of human-associated STEC in Australia. Since 1999, STEC isolates in Australia have been referred to MDU PHL for serotyping and detection of Shiga toxin(s).

A total of 435 human clinical STEC isolates received at MDU PHL between 1 January 2007 and 31 December 2016 underwent WGS and were included in this study. Additional contextual STEC genomes were included, specifically 14 animal STEC isolates received between 2009 and 2013 from bovine or ovine sources, and 36 historical human O157 STEC isolates received from 1993 to 2006 (Supplementary Dataset 1, available in the online version of this article). The genomes were quality-controlled (mean depth of coverage over 50×; mean quality score Q30; assembled genome size after filtering for contigs <500 bp, and genome assembly sizes between 4 800 000 to 5 800 000 bp) and screened *in silico* for the presence of *stx* genes (see below). In total, 485 STEC isolates were included from Australia.

To provide geographical context and enable comparative analysis with STEC in circulation globally, publicly available isolates were included in phylogenetic analysis. In total, 17 representative O157 STEC were included, representing previously reported major lineages of O157 STEC and major clades from the Manning typing scheme [19, 20]. Further, contemporary representatives of other highly prevalent global serogroups [21] circulating globally (O103, O111, O121, O145 and O26) were included, by screening the 500 most recently added isolates to the GenomeTrakr database at the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA319594). Sequences with an SRA accession number were downloaded (access date 6 January 2017), quality-controlled as described above and then screened *in silico* for the presence of *stx* genes (see below). A random selection of ten genomes was included from each serogroup. This resulted in an additional 56 genomes (only six O145 were identified), with three from the UK and 53 from the USA.

Identification and epidemiological typing of the 485 Australian STEC isolates included in this study was performed at MDU PHL. Phenotypic characterization of O- and H-surface antigens was performed using agglutination reactions with rabbit antisera, as previously described [12, 22]. Isolates were determined as STEC by PCR screening for the *stx1* and *stx2* genes using previously described methods [23]. Details of all isolates are given in Supplementary Dataset 1.

### Whole genome sequencing

DNA extraction and WGS was performed at MDU PHL. Sequence libraries were prepared using NexteraXT and sequenced on the NextSeq500 (Illumina NextSeq with 150 bp paired-end reads or MiSeq with 300 bp paired-end reads). Reads are available on the NCBI Sequence Read Archive (BioProject PRJNA319594).

### *In silico* antimicrobial resistance gene detection and multilocus sequence typing

Genome assemblies were constructed using SPAdes (v3.12) using the 'careful' parameter and kmers of 21, 33, 55, 77

[24]. Sequence types (STs) were determined with the programme 'mlst' (https://github.com/tseemann/mlst) using the Achtman database [25]. The genome assemblies of all isolates were also screened for acquired antimicrobial resistance (AMR) determinants in the NCBI resistance database using the programme 'ABRicate' (https://github.com/tseemann/abricateabricate) with a minimum nucleotide identity of 90 % and minimum coverage of 90 %. The results were transformed into a binary dataset in R (version 3.4.0), summarized by clinical drug class and plotted with *ggplot2* (v2.3.0) [26]. Known quinolone resistance mutations (*gyrA* 83 and 87; *parC* 80 and 84) were identified as non-synonymous single nucleotide polymorphisms (SNPs) in these quinolone-resistance determining regions [27].

### *In silico* detection of serotype and *stx* genes

*In silico* O:H serotyping was performed using the EcOH database with SRST2 [28], with the minimum read depth set to 1. The Simpson's Index of Diversity was calculated for both the O-group and the H-types within the most common clonal groups (CGs) using the '*vegan*' package in R (version 3.4.0) [29]. The *stx* genes were initially identified with SRST2 [28], using the same *stx* genes as Ashton *et al.* to construct the toxin database [30]. Mutations in the *stx* genes in CG504 isolates were investigated using ARIBA [31]. *In silico* data were compared to data produced from conventional typing methods (see above). O-groups that fell within previously described O-antigen clusters [32] were classified as concordant. Both A and B subunits had to be detected from the *in silico* screen for the Stx1 or Stx2 complex to be classified as genotypically present.

### Phylogenomic analysis of core genome

WGS reads were trimmed to clip Illumina Nextera adapters and low-quality read data (Phred score <10) using Trimmomatic [33]. The 558 genomes were mapped to the reference chromosome genome O157:H7 Sakai (NCBI accession number: NC_002695.2) using Snippy (v4.3.4)/BWA MEM 0.7.15-r1140 [34] (https://github.com/tseemann/snippy). SNPs were called using Snippy (v4.3.4)/Freebayes requiring a minimum read coverage of 10× at variant sites and minimum proportion of variant evidence of 0.90 [35]. Phage regions were identified in the Sakai reference chromosome using PHASTER [36] and SNPs in these regions were masked. A core genome alignment of 235 746 informative SNPs was produced using snippy-core (v4.3.4) and used as input for IQ-TREE (v1.6.5) [37] with the GTR+G4 model, constant sites (1 590 440, 1 477 994, 1 550 215 and 1 549 891), ultrafast bootstrapping with 1000 replicates, the SH-aLRT parameter with 1000 bootstrap replicates [38] to infer the phylogenetic structure.

Lineages were defined from this tree using RAMI [39] with a patristic distance threshold of 0.0004. This threshold defined two well-characterized lineages within *E. coli,* specifically ST11 (associated with O157:H7), and the precursor to this lineage, ST335 (associated with O55:H7). CGs were identified by the most common ST within the lineage.

### Subset analysis of CG11

In order to further define CG11, the core genome alignment for CG11 was extracted using snippy-core (https://github.com/tseemann/snippy). The resulting SNPs were used to generate a pseudo-whole genome alignment, in which recombination was detected using Gubbins [40], leaving a recombination-filtered core genome alignment of 10 377 SNPs. A recombination-filtered phylogeny of CG11 was first inferred using IQ-TREE (v1.6.5) as described above. We then investigated the temporal signal within CG11 using Tempest (v1.5) [41] by conducting a regression of the root-to-tip branch distances of the CG11 ML phylogeny. The resulting data was visualized in R (version 3.4.0) (Fig. S1a). Pairwise SNP distances between isolates were calculated using the R package *harrietR* (v0.2.3) (https://github.com/andersgs/harrietr). SNP distances between any isolates were filtered to ≤10 using tidyverse (v1.2.1) (https://CRAN.R-project.org/package=tidyverse), before being visualized against the phylogeny using *ggtree* [42]. Putative clusters were defined as groups with two or more isolates with ≤10 pairwise SNPs.

The recombination-filtered SNP alignment was used in BEAST v2.5.1 to estimate a Bayesian phylogeny with divergence dates [43]. The model parameters were GTR+Γ substitution model with alternative clock; strict and relaxed lognormal clock, and alternative population priors; constant coalescent population and constant exponential population tested. The highest supported model was the relaxed lognormal clock under a constant population size and ten independent BEAST runs of 100 million states. These independent runs were combined using LogCombiner v2.5.1, with 20 % burn-in removed and all ESS scores were above 200 and parameter estimates were calculated using Tracer v1.6 [43]. A maximum clade credibility tree was constructed using TreeAnnotator v2.5.1. To test the robustness of the molecular clock signal, five further BEAST runs with randomized tip dates were generated using the highest supported model with 75 million states and burn-in of 20 million [44] (Fig. S1b). The final maximum clade credibility tree was visualized in R using *ggtree* [42].

Clade typing using the Manning scheme [20] was performed within the CG11 group by identifying nine definitive SNPs, eight from Yokoyama *et al.* [45] and the inclusion of additional SNPs [46] from the VCF files produced from mapping analysis produced by Snippy. Primer sequences were used to confirm the correct loci for the eight SNPs in artemis [47], and the reference base confirmed in the O157:H7 Sakai (NCBI accession number: NC_002695.2). Importantly, we observed that some of the loci or bases specified did not match the reference Sakai genome; the nine SNPs used are described in Supplementary Dataset 2.

ISmapper [48] was run with default parameters to screen all read sets for insertion sites of the transposase IS*1203v* (accession AB017524.1) relative to the reference chromosome O157:H7 Sakai (NCBI accession number: NC_002695.2). The binary data was processed in R using *tidyverse* (v1.2.1) (https://CRAN.R-project.org/package=tidyverse). The insertion of a cytosine base in position 125 in *flgF* associated with

a H-non-motile phenotype [49] in Australian CG11 isolates from human cases between 2007–2016 was investigated from the Snippy output by manually identifying insertions at this site.

### Subtyping of *stx* genes in CG11

A combination of mapping and assembly-based approaches were used to infer *stx* subtypes [30] using results from SRST2 and ABRicate, which report calls to the *stx* A and B subunit genes. In isolates with hits to both *stx2A* and *stx2C* alleles, the depth of reads mapping to the *stx2*A subunit from SRST2 was compared to the average depth. Previously subtyped isolates were included in our *stx* subtyping in order to validate our approach [19].

### National surveillance data for STEC

The National Notifiable Disease Surveillance System (NNDSS) was established in Australia in 1990 and coordinates the surveillance of >50 communicable disease or disease groups. Notifications of diseases, such as STEC, are made to the relevant State or Territory health authority, de-identified and then supplied to the Australian Government Department of Health. The NNDSS data for STEC for raw numbers and rates per 100 000 by State/Territory and Year were accessed on 1 March 2018 at . The raw counts and rates per 100 000 for the period of 2007 to 2016 inclusive were plotted in R using *ggplot2* [26].

## RESULTS AND DISCUSSION

### The genomic epidemiology of STEC in Australia

To investigate the population structure of STEC in Australia, we constructed a core genome phylogeny of all 558 STEC genomes included in this study. A total of 52 lineages were identified (based on patristic distance generated with RAMI [39]), of which there were eight CGs with ten or more isolates (Fig. 1). These eight CGs accounted for 85.1 % (475/558) of the STEC genomes included in this study, with the Sakai O157:H7 reference genome also falling within the largest CG, namely CG11. The CG11 lineage, associated with the O157:H7 serotype, accounted for 55.7 % (311/558) of all isolates (Fig. 1). The most common serogroups after O157 in this study were O26 (10.8 %; 59/558 isolates), O111 (7.3 %; 41/558), O128 (3.8 %, 21/558), O-Gp8 (the O117 allele) (1.9 %; 11/558) and O91 (1.8 %; 10/558), with the remaining isolates characterized by over 40 different O-antigens (including 11 for which no *in silico* serotype was determined) (Supplementary Dataset 1), highlighting the diversity of STEC lineages in Australia.

We compared the occurrence of the O157 and several other highly prevalent global serogroups [21] in Australia relative to international isolates, identifying geographic differences between the observed serotypes in the eight CGs. For example, we identified 11 isolates with the O103:H2 serotype (associated with the CG17 lineage), although only one was from Australia (Supplementary Dataset 1). We also identified a rare serotype combination of O103:H25, previously reported as the causative agent of HUS in Norway in 2006, associated with cured mutton sausages [50]. Similar patterns of low representation of Australian STEC in serogroups O121 and O145 was observed. These two serogroups were associated with two STEC lineages, namely CG655 (*n*=11) with O121:H19, and CG32 (*n*=7) with O145:H-. Both lineages were comprised of publicly available isolates from the USA, with the exception of a single Australian isolate in each lineage. Greater numbers of Australian STEC fell within the two highly prevalent global serogroups, O26 and O111. The O26:H11 serotype, associated with CG21, comprised both Australian and international genomes, and the O111 serogroup was identified in serotypes O111:H8, O111:H- (no H-type determined *in silico*) and O111:H11.

The prevalence of acquired AMR determinants was uncommon in the main STEC CGs detected in the dataset (Fig. 2, Supplementary Dataset 3), except for the CG504 lineage, in which multiple AMR determinants were detected, with most isolates in CG504 having acquired AMR genes to between two and five classes of antimicrobials (Figs 2 and S2). No non-synonymous point mutations were detected in *parC,* with only one point mutation in *gyrA* (Asp87Tyr) detected in a O157 historical isolate. Interestingly, in nine isolates a complex mutation in *gyrA*, specifically a deletion at base 248 and an insertion at base 252, resulting in a non-synonymous base change in *gyrA* (Ser83Leu) was detected. Six of these isolates were part of CG504 (Fig. S2). Of note, all isolates within CG504 either had a frameshift mutation (F111fs) or non-synonymous point mutation (L171S) in the *stx1* A subunit; future work should assess the impact of these mutations on Stx1 expression and functionality.

### The O157:H7 STEC lineage II is successfully established in Australia

To investigate the emergence of O157:H7, we used 10 377 SNPs identified in the core genome of the CG11 lineage to construct a timed phylogeny (Fig. 3). Preliminary analysis revealed temporal structure within this lineage (Fig. S1a). The substitution rate (the number of substitutions per site per year) was 3.4E-07 (Fig. S1b), corresponding to previous estimates within *E. coli* pathotypes [19, 51]. Further, the molecular clock signal was robust as no overlap of highest posterior density (HPD) for the substitution rate was observed for the BEAST runs with randomized tip dates (Fig. S1b).

Our data suggest that the most recent common ancestor (MRCA) of all the O157 STEC isolates in our study likely occurred around 1800 (HPD of 1743–1852), consistent with previous findings [19]. We observed geographic variation within the O157 phylogeny, reflected in the distribution of the Australian isolates in the Manning clades (Manning classification; Supplementary Datasets 2 and 4). Notably, only a small proportion of the Australian STEC fell within clade 3, clade 4/5 or clade 6 (all associated with lineage I [19] (Fig. 3). The majority of the Australian isolates (95.1 %, 270/284) in the CG11-specific phylogeny typed as clade 7.

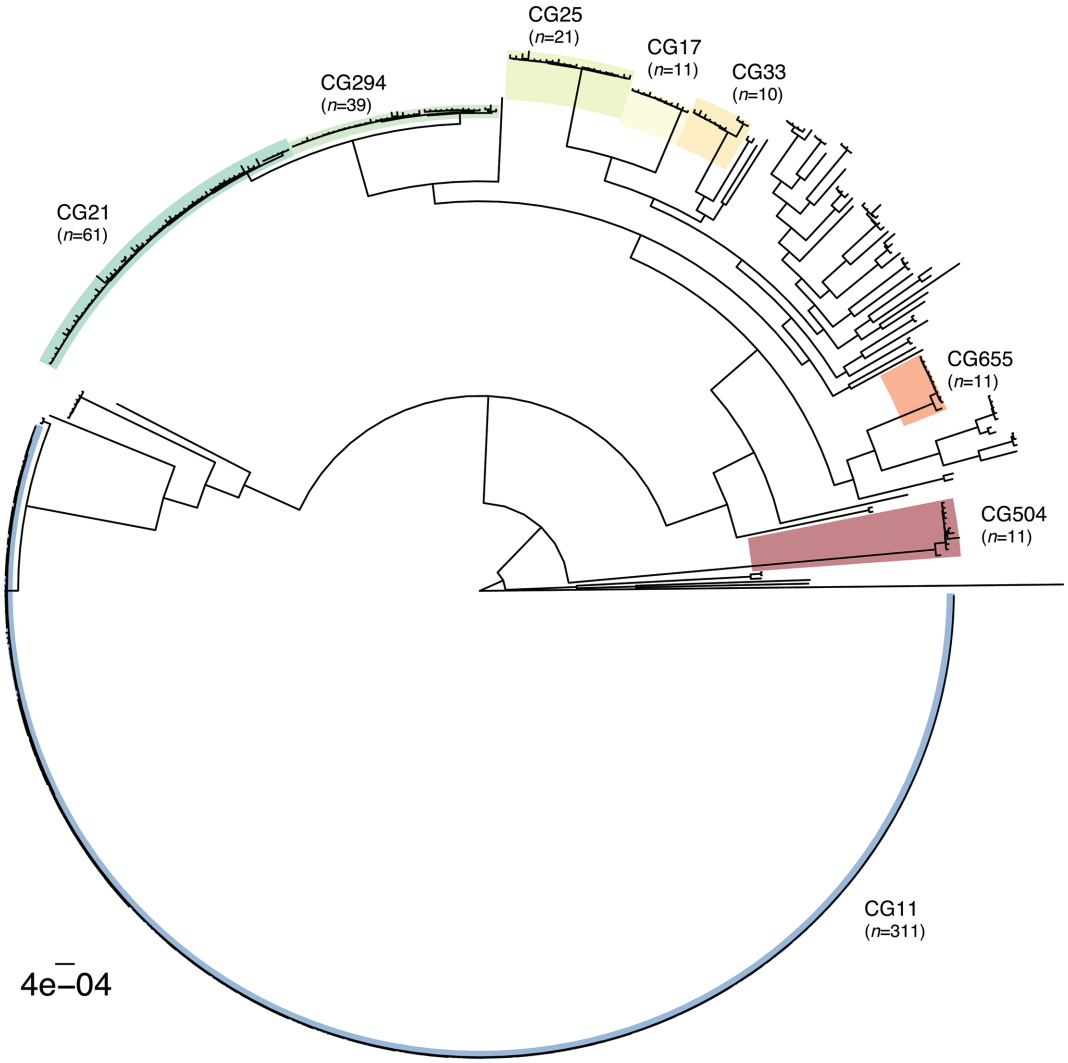


**Fig. 1.** Overview of STEC population structure in Australia. The inferred core phylogeny of 558 STEC genomes produced from 235 746 informative SNPs. The reference genome was O157:H7 Sakai (NCBI accession number: NC_002695.2). Identified CGs are delineated by the most common ST in the CG. The eight major CGs are identified by different colours.

Nearly all human Australian isolates clustered with the lineage II reference isolates [19], as did all ten O157 animal isolates from Australia. Of note, the MRCA for the dominant clade 7 sublineage within Australian 0157 STEC was 1855 (HPD 1830–1877), possibly associated with migration of humans and animal from Europe to Australia. Importantly, the insertion of a cytosine base in position 125 in *flgF* (associated with a non-motile flagellar [49]) was only found in isolates within this sublineage, and was significantly associated with the H-non-motile phenotype (Supplementary Dataset 4). This non-motile phenotype is regarded as a epidemiological marker for Australian O157 STEC. Within Australian 0157 STEC, the most common *stx* subtypes were *stx1a* and *stx2c* (Fig. 3, Supplementary Dataset 4), with interruptions in the *stx2c* A subunit detected in isolates associated with a 2013 outbreak [18]. This genotype has previously been shown to be due to an interruption of the gene by an IS element IS*1203v* [18], and the phylogenetic structure suggests that the outbreak strains emerged locally from the established lineage II (Fig. 3).

Collectively, our data demonstrate that O157 STEC in Australia is distinct from that in the Northern Hemisphere, based upon our comprehensive sampling of Australian STEC over a 10-year period (Fig. S3). Our data support the hypothesis that while there have been multiple introductions of O157 sub-lineages to Australia, the dominant lineage II, associated with the H-non-motile phenotype, has been the most successful in the Australian environment.

## Utility of genomics-based approaches in public health microbiology

Similar to previous studies [10–12], we observed high concordance between O- and H-antigen phenotypic

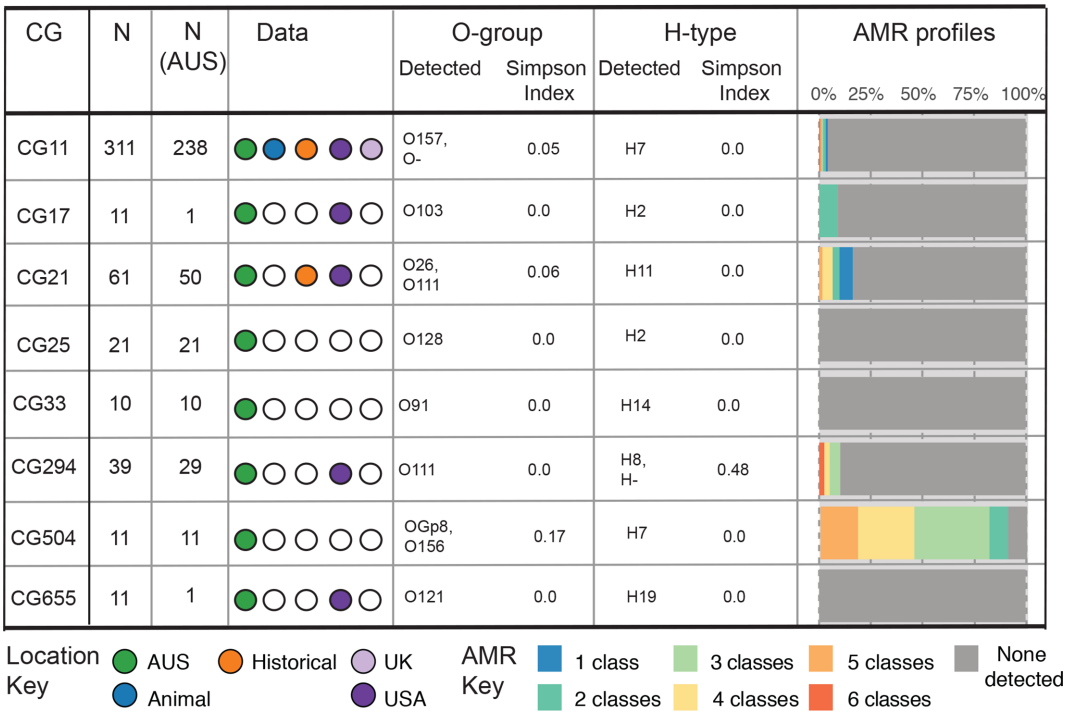


**Fig. 2.** Summary of the features of the eight most common STEC CGs in this study. Data indicates the source (geography or animal or historic Australian) of the genomes included in the CGs. The N (Aus) indicates the number of isolates from humans between 2007–2016. The 'O-group' indicates the distinct O-antigen genes detected in the isolates. The 'H-type' indicates the distinct H-antigen genes detected in the isolates. Where no O-group or H-type was detected *in silico*, this is given by O- and H-, respectively. 'AMR profiles' indicates the proportion of AMR determinants acquired by horizontal gene transfer grouped by drug class detected in the STEC isolates.

serotyping and *in silico* serotyping. *In silico* serotyping detected an O-group in 98.4 % (428/435) of isolates, with 96.4 % concordance (370/384 isolates) between phenotype and genotype (Table S1). *In silico* H-types were detected in 96.1 % (418/435) isolates, and of the 135 isolates with a H-antigen phenotype, concordance was 98.5 % (133/135) (Table S1). Discordance between phenotype and genotype may be due to several factors, including (i) detection of novel antigens (e.g. two isolates with O28 phenotypes characterized with *in silico* OgN9) and (ii) laboratory errors with phenotyping (e.g. transcription or processing errors). Interestingly, of the 300 Hnt isolates in this study, 228 (76 %) were from CG11 and while the *fliC* gene was intact, point mutations in *flgF* were detected (discussed above) (Table S2, Supplementary Dataset 4). The other Hnt isolates were detected in CG33 ($n$=10) and CG294 ($n$=27), with the remaining 35 distributed over 16 different lineages. Future work should explore the presence of additional mutations in the flagellar operon, and assess the presence of these in the Australian setting.

In addition to *in silico* serotyping, we undertook *in silico* determination of the A and B subunit of the *stx* genes, *stx1* and *stx2* respectively, and compared with conventional PCR detection (Table S3). Of the 435 isolates, 78 had *stx1* detected by PCR, 67 had *stx2* and 290 had both *stx1* and *stx2* detected (Supplementary Datasets 1 and 4). Concordance for *stx1* detection (based on *in silico* detection of both subunits) was 97.8 % (360/368), and for *stx2* detection was lower, at 75.1 % (268/357). However, on further investigation, in eight *stx1* and 31 *stx2* 'discordant' isolates, both the A and B subunits were completely absent (using both mapping and BLAST approaches), suggesting a possible loss of Stx phages during long-term storage between the time of PCR detection and WGS. In addition, the *stx2*A subunit gene was incomplete in 54 isolates, with three isolates from 2009 and 51 isolates from 2013 found to have only 84.27 and 48.33 % coverage respectively of this gene. These 54 isolates were detected by conventional PCR as the primer sites remained intact, despite interruption of the gene. All 54 of these isolates clustered in the CG11 lineage (Fig. 3), consistent with a previously reported outbreak [18] that described interruption of the *stx2* A gene by IS*1203v* (GenBank accession AB017524.1). We used ISmapper [48] and Bandage [52] to investigate the interruption of the *stx2* A subunit gene with this IS element in all isolates, but given the fragmented nature of the genomes and the integration of the IS element into another mobile element, we were unable to resolve the interruption of the gene. However, we hypothesize that all of these isolates would have a non-functional form of the Stx2 toxin, as previously suggested [18].

To investigate the potential of genomics-based approaches in identifying putative STEC clusters, we utilized the pairwise SNP distances determined from the recombination-filtered

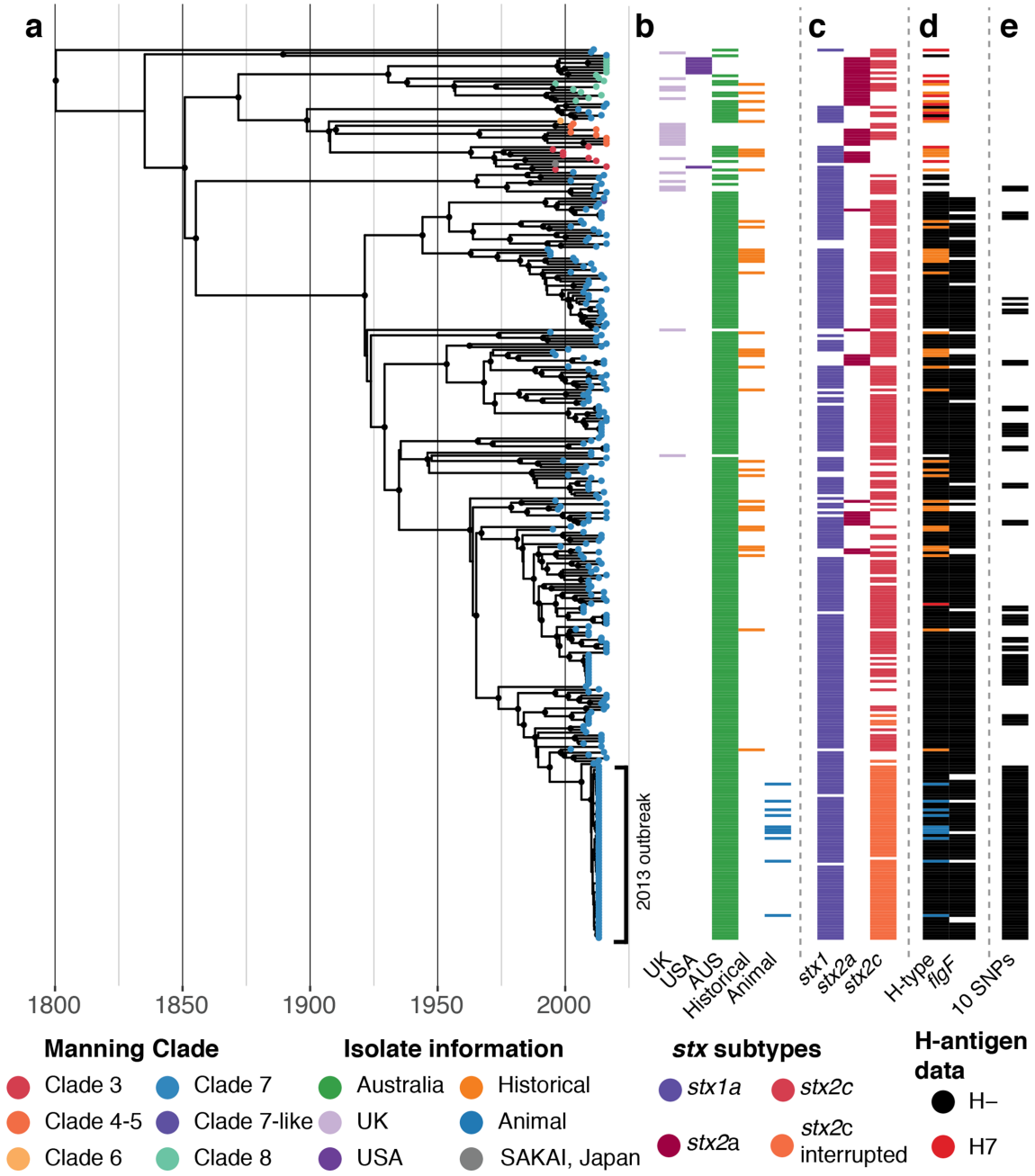


**Fig. 3.** Phylogeny of CG11 suggests emergence of local Australian STEC. (a) Maximum clade credibility phylogeny of 311 isolates belonging to the CG11 lineage, including the reference Sakai genome. Tips are coloured by Manning clade membership [20] determined *in silico*. The scale is in years and nodes of posterior support ≥95 are indicated in the phylogeny by black circles at the node. Isolates from the 2013 STEC outbreak associated with cattle are indicated to the right of the phylogeny [18]. (b) Geographic location of strain collection site (AUS: Australia, UK: United Kingdom and USA: United States of America). (c) The presence of *stx* subtypes is shown to the right of the phylogeny. (d) H-type phenotype data (where available) is shown, together with detection of an insertion of a cytosine base in position 125 in *flgF* associated with a non-motile flagellar phenotype. (e) Isolates that were ten pairwise SNPs from another isolate are shown in black.

CG11 SNP alignment. As an epidemiologically validated 'control' dataset, we used the CG11 isolates from a known outbreak in Queensland, Australia in 2013 [18]. Having established that a single linkage SNP threshold of ≤10 SNPs was able to identify the QLD isolates in the 2013 outbreak, we applied this threshold to other isolates within the CG11 lineage. A total of 18 putative clusters were identified, comprising two or more non-duplicate isolates with a maximum distance of ten SNPs (Supplementary Dataset 5). We identified six additional clusters spanning multiple Australian States, including

an interstate cluster from 2009 from a previously described outbreak [53]. While we did not have detailed epidemiological data to validate these clusters, our study provides a genomic framework for future epidemiological investigations. The detection of known outbreaks, and putative outbreaks spanning jurisdictional boundaries demonstrate the capacity of such an approach to identify geographically dispersed outbreaks of STEC.

Finally, this study also highlights the potential public health impact of culture independent diagnostic testing (CIDT) on genomic surveillance of pathogens. Overall, only 31.3 % of STEC notifications across the study period were associated with a culture (Fig. S3); this proportion dropped further in 2017 due to the increased use of CIDT in one jurisdiction (Fig. S3). Given the potential for STEC to cause large outbreaks, it is critical that concerted efforts are made to ensure continuation of culture-based surveillance.

## Conclusions

Here, we provide insights into the population structure and emergence of STEC in Australia. We demonstrate that the distribution of STEC in Australia differs from STEC in the Northern hemisphere, specifically in that not all of the prevalent global serotypes were highly represented in the Australian STEC collection. Instead, we observed that STEC in Australia comprised a diverse range of lineages, with the most common lineage of O157 STEC (associated with a non-motile phenotype) emerging in the mid-nineteenth century. We further corroborate the public health utility of genomic data for the characterization and surveillance of STEC, and describe a valuable framework for ongoing national genomics-based surveillance of this important public health pathogen.

Funding information
D. A. W. is supported by an NHMRC. Early Career Fellowship (GNT1123854) and B. P. H. is supported by an NHMRC Practitioner Fellowship (GNT1105905). Salary support for D. J. I. was supported by an NHMRC Project Grant (APP1129770) and the European Union's Horizon 2020 research and innovation programme under grant agreement 643476. The Microbiological Diagnostic Unit Public Health Laboratory is funded by the Department of Health and Human Services, Victoria.

Conflicts of interest
The authors declare that there are no conflicts of interest.

Data bibliography
Sequence Read Archive BioProject PRJNA319594 (2018).

References

1. Kirk MD, Pires SM, Black RE, Caipo M, Crump JA *et al*. World health organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS Med* 2015;12:e1001921–1.
2. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M *et al*. Recent advances in understanding enteric pathogenic *Escherichia coli*. *Clin Microbiol Rev* 2013;26:822–880.
3. Ford L, Kirk M, Glass K, Hall G. Sequelae of foodborne illness caused by 5 pathogens, Australia, circa 2010. *Emerg Infect Dis* 2014;20:1865–1871.
4. Lynn RM, O'Brien SJ, Taylor CM, Adak GK, Chart H *et al*. Childhood hemolytic uremic syndrome, United Kingdom and ireland. *Emerg Infect Dis* 2005;11:590–596.
5. Butcher H, Elson R, Chattaway MA, Featherstone CA, Willis C *et al*. Whole genome sequencing improved case ascertainment in an outbreak of shiga toxin-producing *Escherichia coli* O157 associated with raw drinking milk. *Epidemiol Infect* 2016;144:2812–2823.
6. Michino H, Araki K, Minami S, Takaya S, Sakai N *et al*. Massive outbreak of *Escherichia coli* O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *Am J Epidemiol* 1999;150:787–796.
7. Karmali MA. Infection by Shiga toxin-producing *Escherichia coli*: an overview. *Mol Biotechnol* 2004;26:117–122.
8. Frank C, Faber MS, Askar M, Bernard H, Fruth A *et al*. Large and ongoing outbreak of haemolytic uraemic syndrome, Germany, may 2011. *Euro Surveill* 2011;16.
9. Rasko DA, Webster DR, Sahl JW, Bashir A, Boisen N *et al*. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 2011;365:709–717.
10. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT *et al*. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis* 2015;61:305–312.
11. Holmes A, Allison L, Ward M, Dallman TJ, Clark R *et al*. Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. *J Clin Microbiol* 2015;53:3565–3573.
12. Ingle DJ, Holt KE, Levine MM, Kuzevski A, Valcanis M *et al*. *In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb Genom* 2016;2:1–14.
13. Jenkins C, Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. Whole-genome sequencing data for serotyping *Escherichia coli*-It's time for a change! *J Clin Microbiol* 2015;53:2402–2403.
14. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing *Escherichia coli* (STEC) in the United States. *Front Microbiol* 2016;7:457–459.
15. Mellor GE, Fegan N, Gobius KS, Smith HV, Jennison AV *et al*. Geographically distinct *Escherichia coli* O157 isolates differ by lineage, Shiga toxin genotype, and total shiga toxin production. *J Clin Microbiol* 2015;53:579–586.
16. National notifiable disease surveillance system. Available at: http://www9.health.gov.au/cda/source/cda-index.cfm [accessed 1st November, 2018].
17. Motarjemi Y, Moy G, Todd ECD. *Encyclopedia of Food Safety*. Elsevier, Academic Press; 2014.
18. Vasant BR, Stafford RJ, Jennison AV, Bennett SM, Bell RJ *et al*. Mild illness during outbreak of Shiga toxin-producing *Escherichia coli* O157 infections associated with agricultural show, Australia. *Emerg Infect Dis* 2017;23:1686–1689.
19. Dallman TJ, Ashton PM, Byrne L, Perry NT, Petrovska L, Allison L, Gally DL, Wain J *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;1:1–13.
20. Manning SD, Motiwala AS, Springman AC, Qi W, Lacher DW *et al*. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci U S A* 2008;105:4868–4873.
21. Brooks JT, Sowers EG, Wells JG, Greene KD, Griffin PM *et al*. Non-O157 shiga toxin-producing *Escherichia coli* infections in the United States, 1983-2002. *J Infect Dis* 2005;192:1422–1429.
22. Chandler ME, Bettelheim KA. A rapid method of identifying *Escherichia coli* H antigens. *Zentralbl Bakteriol Orig A* 1974;229:74–79.
23. Paton AW, Paton JC. Detection and characterization of Shiga toxigenic *Escherichia coli* by using multiplex PCR assays for *stx1, stx2,*

*eaeA*, enterohemorrhagic *E. coli hlyA*, rfbO111, and rfbO157. *J Clin Microbiol* 1998;36:598–602.

24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M *et al*. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
25. Wirth T, Falush D, Lan R, Colles F, Mensa P *et al*. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* 2006;60:1136–1151.
26. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Cham: Springer International Publishing; 2016.
27. Eaves DJ, Randall L, Gray DT, Buckley A, Woodward MJ *et al*. Prevalence of mutations within the quinolone resistance-determining region of *gyrA*, *gyrB*, *parC*, and *parE* and association with antibiotic resistance in quinolone-resistant *Salmonella enterica*. *Antimicrob Agents Chemother* 2004;48:4012–4015.
28. Inouye M, Dashnow H, Raven L-A, Schultz MB, Pope BJ *et al*. SRST2: rapid genomic surveillance for public health and hospital microbiology Labs. *Genome Med* 2014;6:1–16.
29. OksanenJ, BlanchetFG, KindtR, LegendreP, MinchinMinchin PR, OHaraRB *et al*. 2015. vegan: community ecology package [Internet]. Available from: http://CRAN.R-project.org/package=vegan.
30. Ashton PM, Perry N, Ellis R, Petrovska L, Wain J *et al*. Insight into shiga toxin genes encoded by *Escherichia coli* O157 from whole genome sequencing. *PeerJ* 2015;3:e739–16.
31. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom* 2017;3:1–21.
32. Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K *et al*. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. *DNA Res* 2015;22:101–107.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
34. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26:589–595.
35. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. *arXiv* 2012:1–9.
36. Arndt D, Grant JR, Marcu A, Sajed T, Pon A *et al*. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 2016;44:W16–W21.
37. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
38. Minh BQ, Nguyen MAT, von Haeseler A, Haeseler von A. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol* 2013;30:1188–1195.
39. Pommier T, Canbäck B, Lundberg P, Hagström Åke, Tunlid A. RAMI: a tool for identification and characterization of phylogenetic clusters in microbial communities. *Bioinformatics* 2009;25:736–742.
40. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15–5.
41. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016;2:vew007–7:vew007.
42. Yu G, Smith DK, Zhu H, Guan Y, TT-Y L. ggtree: an rpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2016;8(1:28–36.
43. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H *et al*. Beast 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2014;10:e1003537–6.
44. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J *et al*. Genome-scale rates of evolutionary change in bacteria. *Microb Genom* 2016;2:1–12.
45. Yokoyama E, Hirai S, Hashimoto R, Uchimura M. Clade analysis of enterohemorrhagic *Escherichia coli* serotype O157:H7/H- strains and hierarchy of their phylogenetic relationships. *Infection, Genetics and Evolution* 2012;12:1724–1728.
46. Iyoda S, Manning SD, Seto K, Kimata K, Isobe J *et al*. Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular *stx* subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers. *Open Forum Infect Dis* 2014;1:ofu061.
47. Carver T, Berriman M, Tivey A, Patel C, Böhme U *et al*. Artemis and act: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* 2008;24:2672–2676.
48. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H *et al*. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* 2015;16:1–11.
49. Pintara AP, Guglielmino CJD, Rathnayake IU, Huygens F, Jennison AV. Molecular prediction of the O157:H-negative phenotype prevalent in Australian shiga toxin-producing *Escherichia coli* cases improves concordance of *In Silico* serotyping with phenotypic motility. *J Clin Microbiol* 2018;56:e01906–17–8.
50. Schimmer B, Nygard K, Eriksen H-M, Lassen J, Lindstedt B-A *et al*. Outbreak of haemolytic uraemic syndrome in Norway caused by stx 2-positive *Escherichia* coli O103:H25 traced to cured mutton sausages. *BMC Infect Dis* 2008;8:1073–10.
51. von Mentzer A, Connor TR, Wieler LH, Semmler T, Iguchi A *et al*. Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat Genet* 2014;46:1321–1326.
52. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* 2015;31:3350–3352.
53. Vally H, Hall G, Dyda A, Raupach J, Knope K *et al*. Epidemiology of shiga toxin producing *Escherichia coli* in Australia, 2000-2010. *BMC Public Health* 2012;12:63.