Clinical Infectious Diseases

Whole Genome Sequencing for National Surveillance of Shiga Toxin Producing Escherichia coli O157

--Manuscript Draft--

Manuscript Number:	
Full Title:	Whole Genome Sequencing for National Surveillance of Shiga Toxin Producing Escherichia coli O157
Short Title:	WGS for surveillance of STEC O157
Article Type:	Major Article
Corresponding Author:	Tim Dallman Public Health England Colindale London, UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Public Health England Colindale
Corresponding Author's Secondary Institution:	
First Author:	Tim Dallman
First Author Secondary Information:	
Order of Authors:	Tim Dallman
	Lisa Byrne
	Philip Ashton
	Lauren Cowley
	Neil Perry
	Goutam Adak
	Liljana Petrovska
	Richard Ellis
	Richard Elson
	Anthony Underwood
	Jonathan Green
	William Hanage
	Claire Jenkins
	Kathie Grant
	John Wain
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Background National surveillance of gastrointestinal pathogens, such as Shiga toxin-producing Escherichia coli O157 (STEC O157), is key to rapidly identifying linked cases in the distributed food network to facilitate public health interventions. In this study we use whole-genome sequencing (WGS) as a tool to inform national surveillance of STEC O157 in terms of identifying linked cases, clusters and guiding epidemiological investigation. Methods We retrospectively analysed 334 isolates randomly sampled from 1002 strains of

	STEC O157 received by the Gastrointestinal Bacteria Reference Unit (GBRU) at Public Health England, Colindale in 2012. The genetic distance between each isolate, as estimated by WGS, was calculated and phylogenetic methods used to place strains in an evolutionary context. Results Estimates of linked clusters representing STEC outbreaks in England and Wales increased by two fold when WGS was used instead of traditional typing techniques. The previously unidentified clusters were often widely geographically distributed and small in size. Phylogenetic analysis facilitated identification of temporally distinct cases sharing common exposures and delineating those that shared epidemiological and temporal links. Comparison with Multilocus Variable Analysis (MLVA) showed that while MLVA is as sensitive as WGS, WGS provides a more timely resolution to outbreak clustering. Conclusions WGS has come of age as a molecular typing tool to inform national surveillance of STEC O157; it can be used in real-time to provide the highest strain level resolution for outbreak investigation. WGS allows linked cases to be identified with unprecedented specificity and sensitivity that will facilitate targeted and appropriate public health investigations.
Suggested Reviewers:	Alfredo Caprioli alfredo.caprioli@iss.it Head of STEC WHO Reference Laboratory
	Flemming Scheutz FSC@ssi.dk WHO Collaborating Centre for Reference and Research on Escherichia and Klebsiella
	David Gally david.gally@roslin.ed.ac.uk Professor at Roslin Institute - renowned world leader on STEC O157 research
	Nicholas Thomson nicholas.thomson@sanger.ac.uk Professor Nicholas Thomson's research uses genomic techniques to understand infectious disease in the context of global health.
Opposed Reviewers:	

Dear Sir

We would like to submit our research article ‘Whole genome sequencing for National Surveillance of Shiga Toxin Producing Escherichia coli O157’ for your consideration. Whole genome sequencing has the potential to revolutionise outbreak investigation and infectious disease surveillance by providing unprecedented strain level resolution. Recently, a number of excellent articles have been published in your journal on the impact of WGS on outbreak investigations. In our study, we show the impact of WGS on national surveillance for the first time using the clinically important gastrointestinal pathogen, STEC O157, as an exemplar.

STEC O157 can be transmitted via food and water, direct contact with animals or contact with a contaminated environment and, therefore, determining the source of outbreaks is challenging. Current typing methods have relatively low strain discrimination or are labour-intensive and time-consuming. Therefore, for STEC O157 surveillance, WGS represents the ultimate in typing; rapidly identifying linked cases with unprecedented sensitivity and specificity. We show that by routinely sequencing STEC O157, twice as many clusters of STEC O157 will be identified in England, compared to the number identified using current methods. The detection of foodborne outbreaks that are currently occurring “under the radar” will have a major impact on food safety interventions and public health policy.

We believe that routine WGS characterisation of dispersed infectious disease, such as STEC O157 will give rise to a paradigm shift in how public health centres perform national surveillance as linked cases will be identified on genomic similarity alone. This will rapidly inform appropriate and targeted public health investigations. WGS presents an effective method to monitor the movement of pathogens in the global food distribution networks.

We believe this research article is pertinent to the Clinical Infectious Disease journal as the adoption of this methodology for national and global surveillance will have such a dramatic impact in public health surveillance practices. Thank you for your time and we look forward to receiving your feedback.

Yours sincerely,

Dr Tim Dallman

Whole Genome Sequencing for National Surveillance of Shiga Toxin Producing *Escherichia coli* O157

Authors:

Timothy J Dallman[1#], Lisa Byrne[1#], Philip M Ashton[1], Lauren A Cowley[1], Neil T Perry[1],Goutam Adak[1], Liljana Petrovska[4]Richard J Ellis[4], Richard Elson[1], Anthony Underwood[1], Jonathan Green[1], William P Hanage[2], Claire Jenkins*[1], Kathie Grant[1], John Wain[3]

*** Corresponding author**

[#] These authors contributed equally

**Public Health England[1]
61 Colindale Avenue
London
NW9 5EQ**

**Harvard School of Public Health[2]
Huntington Avenue
Boston
MA**

**University of East Anglia[3]
Norwich
NR4 7TJ**

**Animal Health and Veterinary Laboratories Agency[4]
Woodham Lane
Surrey
KT15 3NB**

Summary

Background

National surveillance of gastrointestinal pathogens, such as Shiga toxin-producing *Escherichia coli* O157 (STEC O157), is key to rapidly identifying linked cases in the distributed food network to facilitate public health interventions. In this study we assess the use of whole-genome sequencing (WGS) as a tool to inform national surveillance of STEC O157 in terms of identifying linked cases, clusters and guiding epidemiological investigation.

Methods

We retrospectively analysed 334 isolates randomly sampled from 1002 strains of STEC O157 received by the Gastrointestinal Bacteria Reference Unit (GBRU) at Public Health England, Colindale in 2012. The genetic distance between each isolate of STEC O157 as estimated by WGS was calculated and phylogenetic methods used to place strains in an evolutionary context.

Findings

Estimates of linked clusters representing STEC outbreaks in England and Wales increased by two fold when WGS was used instead of traditional typing techniques. The previously unidentified clusters were often widely geographically distributed and small in size. Phylogenetic analysis facilitated the identification of temporally distinct cases that shared common exposures as well as delineating those that shared epidemiological and temporal links. Comparison with MLVA the current gold standard molecular epidemiology tool showed that while MLVA is as sensitive in linking cases the method fails to resolve clusters as timely as WGS.

Interpretation

WGS has come of age as a molecular typing tool to inform national surveillance of STEC O157; it can be used in real-time to provide the highest strain level resolution information for outbreak investigation. WGS will allow linked cases to be identified with unprecedented specificity and sensitivity that will facilitate targeted and appropriate public health investigations.

Funding

National Institute for Health Research scientific research development fund (108601)

Introduction

Gastrointestinal disease is an important public health problem in England with up to 20% of the population experiencing at least one episode of acute gastroenteritis each year [1]. An effective national surveillance program for gastrointestinal diseases is imperative to identify cases with linked exposures; this is especially pertinent for pathogens which may enter nationally distributed food networks. Whilst conventional epidemiological investigation using detailed questionnaires and contact tracing is vital, to achieve optimal surveillance we must complement these activities with a rapid and robust molecular typing method to accurately discriminate between linked cases and sporadic infections.

With over 1000 presumptive isolates submitted to the Gastrointestinal Bacteria Reference Unit (GBRU) annually [2] infections with Shiga toxin-producing *Escherichia coli* O157 (STEC O157) continue to exert a public health burden in England, both economically and in terms of morbidity and mortality. Symptoms of STEC infections range from mild to severe but typically include bloody diarrhoea. Approximately 6% of cases develop haemolytic uraemic syndrome [3] a disease which affects the blood and kidneys and most frequently affects children. In some cases the disease can be fatal.

The main reservoir of STEC in England is cattle, although it is carried by other animals, mainly ruminants [4, 5]. Transmission to humans occurs through direct or indirect contact with animals or their environments; consumption of contaminated food or water, and through person-to-person contact [6-8]. Contamination of the food-supply can cause large-scale national and multinational outbreaks [9-11].

Outbreaks, involving two or more cases in different households or residential institutions, vary in number annually but since 2009 have contributed between 9% and 25% of isolates in England and Wales (GBRU/ Department of Gastrointestinal Emerging and Zoonotic Infections (GEZI) (in-house data) with the majority of cases occurring within households or are, apparently, sporadic. All isolates received by GBRU are routinely phage typed [12], but in England, the majority (60%) of isolates are either PT8 or PT21/28, and so the ability of this method to discriminate between cases resulting from separate exposures is very low (GBRU in-house data) leading to the hypothesis that additional "outbreaks" are occurring under the surveillance radar. Multi Locus Variable Number Tandem Repeat Analysis (MLVA) has previously been used reactively when an exceedance of a particular phage type has been detected and is now been used in real-time by GBRU.

The utility of whole genome sequencing for the investigation of outbreaks has already been demonstrated for several bacterial pathogens [13, 14] and there is increasing evidence in the literature for the positive contribution of WGS to outbreaks involving gastrointestinal pathogens [15-19] . The aim of this study was to expand the use of WGS by evaluating for the first time a whole genome sequencing approach to inform national surveillance of a major pathogen. Firstly, by validating the WGS approach using clearly defined outbreak and sporadic cases of STEC

O157 and, secondly, by investigating the insights WGS can provide additional insights into outbreak definition, transmission networks, and other aspects of the underlying epidemiology of this pathogen.

Methods

Strain selection

A total of 425 isolates were selected for sequencing; 334 isolates were randomly selected from 1002 STEC O157 culture positive isolates received by GBRU from cases in England, Wales and Northern Ireland during 2012. An additional 91 English historical isolates received between 1990 and 2011 were selected to provide context as a sample of the background population. The total collection contained strains from known outbreaks, household clusters, serial strains isolated from the same patient and strains from apparently sporadic cases. A total of 18 phage types [20] were represented.

Genome Sequencing and Sequence Analysis

Genomic DNA was fragmented and tagged for multiplexing with Nextera XT DNA Sample Preparation Kits (Illumina) and sequenced at the AHVLA using the Illumina GAII platform with 2x150bp reads. Short reads were mapped to the reference STEC O157 strain *Sakai* [21] using BWA-SW[22]. The Sequence Alignment Map output from BWA was sorted and indexed to produce a Binary Alignment Map (BAM) using Samtools [23]. GATK2 [24] was used to create a Variant Call Format (VCF) file from each of the BAMs, which were further parsed to extract only single nucleotide polymorphism (SNP) positions which were of high quality in all genomes (MQ>30, DP>10, GQ>30, Variant Ratio >0.9). Pseudosequences of polymorphic positions were used to create approximate maximum likelihood trees using FastTree [25] under the Jukes-Cantor model of nucleotide evolution. Pair-wise SNPS distances between each pseudosequence (normalised by size of the shared core genome) were calculated. FASTQ sequences were deposited in the NCBI Short Read Archive under the bioproject PRJNA248042.

Data Handling

Local Laboratories reported presumptive isolates of STEC directly to PHE Centres who arrange for STEC Enhanced Surveillance Questionnaires (SESQ) to be administered to patients. The SESQ collects demographic details; risk status; clinical condition (including progression to HUS); household or other close contact details; exposures including travel, food and water consumption, contact with animals and environmental factors; epidemiological case classification; and outbreak /cluster status. Completed questionnaires are forwarded for inclusion in the National Enhanced Surveillance System for STEC (NESSS) in England which is managed by GEZI. SESQ data were reviewed for each selected strain and strains classified in respect to known outbreak status, known household cluster status or whether multiple isolates originated from the same patient. Any strains fulfilling these criteria were designated as having a known epidemiological link.

Pair-wise SNP distances were calculated for all strains in this study. In previously reported outbreaks, onset of illness in cases occur a median of 39 days from another linked case with a mode of 1 (in-house data).Using specimen dates of isolates, temporality between isolates of different genetic distances were compared. The pair-wise

SNP distribution and temporal links between known linked cases was examined and a relatedness threshold determined accordingly. As related strains are likely to originate from a common source the threshold was termed the Common Source Threshold (CST). This threshold was then applied to all other strains in the dataset and evaluated for epidemiological context.

Related strains within the CST were classified into clusters on the basis of having at least one SNP distance within the CST to another isolate in the dataset. Clusters not previously identified were designated WGS linked clusters. Temporal and geographic links between cases in clusters were examined and comparisons made between epidemiologically identified and WGS linked clusters.

Deeper phylogenetic relationships were also investigated to ascertain whether they provided epidemiologically useful information or associations. Clusters of 25 SNP genetic distance were constructed (herby referred to as phylogenetic clusters (PCs) and those with more than one CST cluster within each PC were investigated for shared epidemiological associations.

All STEC O157 isolates reported between 1 May 2012 and 31 December 2013 which have been both typed through MLVA and whole genome sequenced were used to investigate clustering dynamics for each method. Survival analysis was used to test the null hypothesis that there is no difference in timeliness and completeness of clustering related isolates using the two methods. For survival analysis, an isolate clustering with another isolate based on <=1 Locus Variant for MLVA or <CST for WGS represented a failure. Across the study period, isolates will enter at various time points based on laboratory report date. At that point the isolate is at risk of clustering with other isolates already in the study population or isolates entering the study at a later date. Kaplan-Meier estimates of the survivor function was estimated for both methods and displayed as cumulative survival curves with accompanying tables present those at risk at specific time points. The proportional hazards assumption (PHA) was tested by plotting the log cumulative hazard in both groups, where the PHA applied, the survival function in the two groups was compared by calculating a hazard ratio using Cox regression.

Results

Distribution of pair-wise distance between closely related isolates

For 183 out of 425 strains used in this study an epidemiological link to at least one other case was known. This included 16 where multiple isolates were sequenced from the same person, 43 isolates part of 26 separate household clusters, and 124 cases part of 14 known outbreaks. The remaining 242 strains had no common link previously identified. The pair-wise SNP distance distribution revealed that no pair of epidemiologically linked isolates had greater than 5 SNP differences with a mean of 1 SNP in isolates from same household (SD=0·99) or known common source (SD=1·04) and 0·3 SNPs (SD=0·60) from isolates from the same person (see Figure 1).

There were 136 cases with no known epidemiological link that were within 5 or less SNPs to another case. The majority (87%) of pairs that fell within the 5 SNP threshold, comprised strains isolated within 30 days of each other with a mean interval between pairs of samples being 11 days. Between a genetic distance of 5-10 SNPS the mean interval between pairs of samples increased to 258 days (Figure 2). As all previously linked isolates fell within a 5 SNP threshold and the majority of pairs of cases within this threshold were temporally linked we hypothesis a threshold of 5 SNP to categorise isolates as related. As related in this context alludes to a common source of infection, strains that are within 5 SNPs of another are referred to as within the Common Source Threshold (CST).

Applying the Common Source Threshold

160 strains isolated during 2012 fell within the CST. These strains can be formed into 53 clusters where members of the cluster must share at least one link within the CST. Twenty of the clusters (46 strains) represented either household outbreaks or multiple strains from the same patient. The remaining 33 of 114 strains represented 34% of the dataset. Routine public health investigation previously undertaken had not identified 20/33 clusters and were designated “WGS linked" clusters. Of the 20 WGS linked clusters, 18 comprised between two and four cases, while two larger clusters comprised 12 and 7 cases. Overall, if we conclude that all cases within the CST are part of epidemiologically linked clusters this corresponds to an increase in sensitivity of greater than 50% in detecting linked cases outside the household setting when using whole genome sequencing to supplement the current approach.

Epidemiology of WGS linked clusters

The 20 WGS linked clusters were statistically more geographically dispersed than the 13 epidemiologically linked clusters (Figure 3a) with a mean residential distance of 169km (standard deviation, 111km) for the former and 29km (standard deviation, 34km) for the later ($p=6{\cdot}0e^{-5}$ one tailed T-Test). Strains of STEC O157 associated with a large national foodborne PT8 outbreak from 2011 [9] and a petting farm PT21/28 outbreak [26] were included for context (Figure 3b). The geographical dispersal of cases linked by WGS mirrors the distribution of the national PT8 outbreak as well as encompassing the distribution of a geographically restricted outbreak. Conversely the epidemiologically linked clusters most closely mirrored the geographically restricted outbreak highlighting the difficulty in recognising national distributed cases without high resolution strain discrimination such as WGS.

Retrospective epidemiological follow-up was undertaken for cases in the two larger clusters. One cluster comprised 12 nationally distributed cases with onset dates all within 15 days of each other. No common exposure factors were identified through review of the SESQs, however the epidemic curve and national distribution of cases was indicative of a food-borne source of infection. Nine cases were re-interviewed using a bespoke follow-up questionnaire focusing on food consumption. The only common exposure among reported was the consumption of a specific pre-packed foodstuff from different branches of one major supermarket chain. Due to the limited number of cases, it was not possible to undertake further analytical epidemiology.

The second largest cluster contained seven cases. Four cases were from separate English public health regions with onset dates spanning a two-week period. SESQs were again reviewed and it was identified that 3 cases had visited the same village, where another case was resident, within the incubation period. These four cases reported visiting the same public house within a three day period but shared no common foodborne exposure. All four cases had engaged in recreational activities (e.g. walking in a national park) putting them at risk of environmental exposure. Three additional cases in this cluster (0 SNP difference) later reported onset dates between three and four weeks after the first four cases. These three cases came from different regions, did not report travel to the same location as the first four and shared no obvious exposures suggesting the cases were exposed to the same source of infection but via different routes and/or vehicles.

Outbreak detection MLVA vs WGS

Clustering based on the WGS defined common source threshold increased sensitivity in identifying linked cases, however it is also necessary to compare this approach to other fine-typing methods deployed for STEC O157, e.g. MLVA. Using a survival analysis of X samples typed by both methods in 2012 survival (i.e. not clustering with another isolate) showed no significant difference with MLVA vs WGS CST based on clustering a single isolate with another (Log rank test for equality of survivor function: p=0.101 Cox Hazard Ratio=0.89, p=0.198) (Figure 4). This indicates there is no difference in timeliness of clustering between the two methods. However, when we consider the time to cluster completion from the cluster event this is a significant speed increase in time to completion of clusters with WGS CST apposed to MLVA (Log rank test for equality of survivor function: p=0.0006, Cox Hazard Ratio=1.44, p=0.001) (Figure 5).

Epidemiological Context of Phylogenetic Clusters

Cases within the CST represent temporally linked cases and these have been shown to include cases with common epidemiological exposures. Although the temporal relationship between pairs quickly dissipated as the genetic distance moved outside the CST, we investigated whether deeper phylogenetic relationships also provided epidemiologically useful information or associations. Nineteen PCs (see Methods) were identified, and 10 had no geographical association or common exposures between the CST clusters within as assessed through the SESQ. One PC contained 3 CST clusters sharing a common exposure to a national park in the midlands (Figure 6). The different CST clusters within this PC correlated with year of isolation highlighting the potential to identify the persistence of strains in the environment over time. Conversely, several temporally related isolates associated with this national park fall into two separate CST clusters, separated phylogenetically with a non-temporally related strain, highlighting the ability of WGS to delineate closely related circulating strains from the same environmental source.

Two PCs contained CST clusters where the majority of strains were of Northern Irish provenance. Those cases that were not resident in N. Ireland reported travel to various parts of the province in their enhanced surveillance questionnaire. Similarly, PCs were identified with cases associated with Wales and travel to the Middle East. Figure 7 shows the distribution of PC's and CST's on a maximum likelihood phylogeny.

Discussion

In this study, the potential of WGS in national surveillance of STEC O157 was assessed for its ability to improve outbreak detection, and provide additional insights over conventional epidemiological investigations. WGS confirmed that strains from the same patient, from cases within the same household and from cases with known epidemiological links had little or no difference in their core genomes. These cases fell within a 5 SNP threshold within which we found strong temporal correlations suggestive of epidemiological linkage. Using this empirically observed cut-off of 5 SNPs we could determine with unprecedented clarity which strains of STEC O157 were likely to be epidemiologically linked. WGS detected linked cases of STEC O157 in 334 representative strains from an annual season with twice the sensitivity of current methods. This suggests that current outbreak detection is highly specific, but comparatively insensitive, and that the previous estimate of outbreaks, involving two or more cases in different households or residential institutions, contributing between 9% and 25% of isolates in England and Wales is conservative. Previously elusive clusters were often more geographically dispersed than those identified using the traditional approach. It is suggested that these geographically dispersed outbreaks with no obvious common exposures are foodborne. This type of outbreak profiling will facilitate outbreak investigations through focusing hypothesis generation on food-borne exposures at an early stage.

Two previously unidentified clusters were re-examined in light of the WGS sequence analysis. One cluster showed common exposure to a specific foodstuff from one super-market chain. Another cluster showed common exposure in several cases to a public house in the NW of England although later cases within the cluster did not share this exposure indicating a possible different route of transmission.

MLVA is the current gold standard fine-typing method for STEC O157. In this study we show that for identifying linked cases the current thresholds of one locus variant or less for clustering provides the same sensitivity as using the WGS CST. This is an important finding as it not only gives confidence in the interpretation of MLVA to those public health laboratories not yet ready to adopt WGS methodologies but also allows cross communication of results between practitioners of these two techniques. An important distinction between the two methods is the time it takes to resolve the complete clusters of cases within an outbreak with WGS CST completing clusters significantly faster than MLVA. This feature can by explained by the fact all linked cases tend to fall within the CST for all cases where as in a large MLVA cluster several isolates will only be joined via an intermediately isolate i.e. double locus variants joined by a shared single locus variant. This phenomena has implications in accurately defining the microbiological case definition at the start of an outbreak investigation as outbreaks that resolve themselves to a single cluster may appear as multiple clusters until intermediate isolates are sampled.

The phylogenetic context of common source clusters was analysed to see if there was any epidemiological signal between separate but related common source events. Several regional or travel associated PCs were identified highlighting the geographical isolation of STEC O157 even within the British Isles. The geographical signal observed in the WGS of STEC O157 has been described previously [27] and has obvious implications in facilitating outbreak investigations. For example, isolates could be linked to food sourced from specific regions of the world or cases could be ruled out of a point source outbreak by confirming their strain originated from further afield, given adequate sampling of potential source populations. This highlights the potential of WGS to not only identify linked cases with high sensitivity and accuracy but also to provide long term epidemiological context through strains that are phylogenetically related

A PC was also identified where the close contact clusters represented temporally distinct cases that shared a common exposure to a national park. This highlights the potential to identify recurrence of infection from a common environmental source over time as well as persistence of a strain over a number of years. Within this cluster, WGS could also discriminate between closely related temporally conserved strains highlighting different exposures of related strains from the same environment.

The primary aims of gastrointestinal disease surveillance are to identify outbreaks, monitor long term trends and inform the effectiveness of policy and other public health interventions. These results show that the impact of whole genome sequencing of STEC O157 on national surveillance is considerable. Clusters of infection provide windows of opportunity for investigation and WGS demonstrates unparalleled sensitivity and accuracy in identifying linked cases coupled with phylogenetic clustering of how strains are related over time and space. At the same time, its ability to accurately define sporadic cases over time enables better characterization of the population at risk and to assess the relative importance of exposures leading to these infections, which may differ from those leading to outbreaks.

Timely analysis and interpretation of WGS data will inform public health interventions by identifying linked cases (i.e. early warning of outbreaks) as well as inferring epidemiological context through evolutionary relationships. Furthermore the ability to unambiguously rule out associations will prevent inappropriate public health actions from being taken saving resources at the health protection and local authority level. Whilst this study highlights the impact based on the sequencing of clinical isolates, for the true potential of the WGS approach to be realised, parallel efforts need to be initiated in the agriculture, veterinary and food industries. Good communication and rapid sharing of real-time STEC WGS data with colleagues working across these industries will allow evidence-based trace back of isolates to their source and reveal specific risk factors in the food chain and environment, thus facilitating the targeting of resources and public health interventions in order to have maximum impact on reducing the burden of STEC O157 disease in England.

Acknowledgments

Reference List

1. Wheeler JG, Sethi D, Cowden JM, Wall PG, Rodrigues LC, Tompkins DS, Hudson MJ, Roderick PJ. Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. BMJ 1999;318:1046-1050.

2. Jenkins C, Lawson A, Cheasty T, Bolton E, Smith G. Assessment of a real-time PCR for the detection and characterisation of Verocytotoxigenic Escherichia coli. J Med Microbiol 2012;%19..

3. Lynn RM, O'Brien SJ, Taylor CM, Adak GK, Chart H, Cheasty T, Coia JE, Gillespie IA, Locking ME, Reilly WJ, Smith HR, Waters A, Willshaw GA. Childhood hemolytic uremic syndrome, United Kingdom and Ireland. Emerg Infect Dis 2005;11:590-596.

4. Pennington H. Escherichia coli O157. Lancet 2010;376:1428-1435.

5. Ferens WA, Hovde CJ. Escherichia coli O157:H7: animal reservoir and sources of human infection. Foodborne Pathog Dis 2011;8:465-487.

6. Locking ME, O'Brien SJ, Reilly WJ, Wright EM, Campbell DM, Coia JE, Browning LM, Ramsay CN. Risk factors for sporadic cases of Escherichia coli O157 infection: the importance of contact with animal excreta. Epidemiol Infect 2001;127:215-220.

7. Gillespie IA, O'Brien SJ, Adak GK, Cheasty T, Willshaw G. Foodborne general outbreaks of Shiga toxin-producing Escherichia coli O157 in England and Wales 1992-2002: where are the risks? Epidemiol Infect 2005;133:803-808.

8. Pritchard GC, Smith R, Ellis-Iversen J, Cheasty T, Willshaw GA. Verocytotoxigenic Escherichia coli O157 in animals on public amenity premises in England and Wales, 1997 to 2007. Vet Rec 2009;164:545-549.

9. Perry N, Cheasty T, Dallman T, Launders N, Willshaw G. Application of multi-locus variable number tandem repeat analysis to monitor Verocytotoxin-producing Escherichia coli O157 phage type 8 in England and Wales: emergence of a profile associated with a national outbreak. J Appl Microbiol 2013;10.

10. Buchholz U, Bernard H, Werber D, Bohmer MM, Remschmidt C, Wilking H, Delere Y, an der HM, Adlhoch C, Dreesman J, Ehlers J, Ethelberg S, Faber M, Frank C, Fricke G, Greiner M, Hohle M, Ivarsson S, Jark U, Kirchner M, Koch J, Krause G, Luber P, Rosner B, Stark K, Kuhne M. German outbreak of Escherichia coli O104:H4 associated with sprouts. N Engl J Med 2011;365:1763-1770.

11. Bell BP, Goldoft M, Griffin PM, Davis MA, Gordon DC, Tarr PI, Bartleson CA, Lewis JH, Barrett TJ, Wells JG, . A multistate outbreak of Escherichia coli O157:H7-associated bloody diarrhea and hemolytic uremic syndrome from hamburgers. The Washington experience. JAMA 1994;272:1349-1353.

12. Khakhria R, Duck D, Lior H. Extended phage-typing scheme for Escherichia coli O157:H7. Epidemiol Infect 1990;105:511-520.

13. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 2013;13:137-146.

14. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, Piazza P, Wilson DJ, Bowden R, Donnelly PJ, Dingle KE, Wilcox M, Walker AS, Crook DW, TE AP, Harding RM. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol 2012;13:R118.

15. Gilmour MW, Graham M, Van DG, Tyler S, Kent H, Trout-Yakel KM, Larios O, Allen V, Lee B, Nadon C. High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. BMC Genomics 2010;11:120.:120.

16. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W, McLaughlin SF, Henkhaus JK, Leopold B, Bielaszewska M, Prager R, Brzoska PM, Moore RL, Guenther S, Rothberg JM, Karch H. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 2011;6:e22751.

17. Underwood AP, Dallman T, Thomson NR, Williams M, Harker K, Perry N, Adak B, Willshaw G, Cheasty T, Green J, Dougan G, Parkhill J, Wain J. Public health value of next-generation DNA sequencing of enterohemorrhagic Escherichia coli isolates from an outbreak. J Clin Microbiol 2013;51:232-237.

18. McDonnell J, Dallman T, Atkin S, Turbitt DA, Connor TR, Grant KA, Thomson NR, Jenkins C. Retrospective analysis of whole genome sequencing compared to prospective typing data in further informing the epidemiological investigation of an outbreak of Shigella sonnei in the UK. Epidemiol Infect 2013;1-8.

19. Allard MW, Luo Y, Strain E, Li C, Keys CE, Son I, Stones R, Musser SM, Brown EW. High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach. BMC Genomics 2012;%19;13:32. doi: 10.1186/1471-2164-13-32.:32-13.

20. Willshaw GA, Smith HR, Cheasty T, Wall PG, Rowe B. Vero cytotoxin-producing Escherichia coli O157 outbreaks in England and Wales, 1995: phenotypic methods and genotypic subtyping. Emerg Infect Dis 1997;3:561-565.

21. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001;8:11-22.

22. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-595.

23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-2079.

24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-1303.

25. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 2010;5:e9490.

26. Ihekweazu C, Carroll K, Adak B, Smith G, Pritchard GC, Gillespie IA, Verlander NQ, Harvey-Vince L, Reacher M, Edeghere O, Sultan B, Cooper R, Morgan G, Kinross PT, Boxall NS, Iversen A, Bickler G. Large outbreak of verocytotoxin-producing Escherichia coli O157 infection in visitors to a petting farm in South East England, 2009. Epidemiol Infect 2011;1-14.

27. Mellor GE, Sim EM, Barlow RS, D'Astek BA, Galli L, Chinen I, Rivas M, Gobius KS. 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 2012;78:4724-4731.

Figure 1. Histogram showing proportion of pairs against SNP distance of cases with a known epidemiological link.

Figure 2. Histogram showing frequency of pairs against SNP distance. Each bar is coloured as a proportion of pairs isolated within 7 days, 7 to 30 days and greater than 30 days.

Figure 3a. Scatter diagram showing the average pairwise residential distance of each close contact cluster against the size in number of cases. The colouring represents whether the cluster was already identified through epidemiological investigation or if identified by whole genome sequencing alone. Figure 3b. Histogram showing the distribution of residential distance for WGS linked clusters and epidemiologically linked clusters. PT8 National and PT21/28 Farm represent distributed food-borne and point source outbreaks respectively.

Figure 4. Maximum likelihood phylogeny of 15 isolates associated with cases that visiting the same national park. The clusters represent 3 different common source threshold clusters within a single phylogenetic cluster. The level of resolution allows the delineation of strains from different years. The strain in red was temporally related to the strains in blue but significantly different gnomically to suggest a different source of STEC exposure.

Figure 5. Maximum likelihood phylogeny of X isolates. Common source threshold clusters identified through WGS alone are coloured red and those identified through traditional methods coloured blue. Phylogenetic clusters that contained strains with related exposures are shaded green.

Figure 1
Click here to download high resolution image

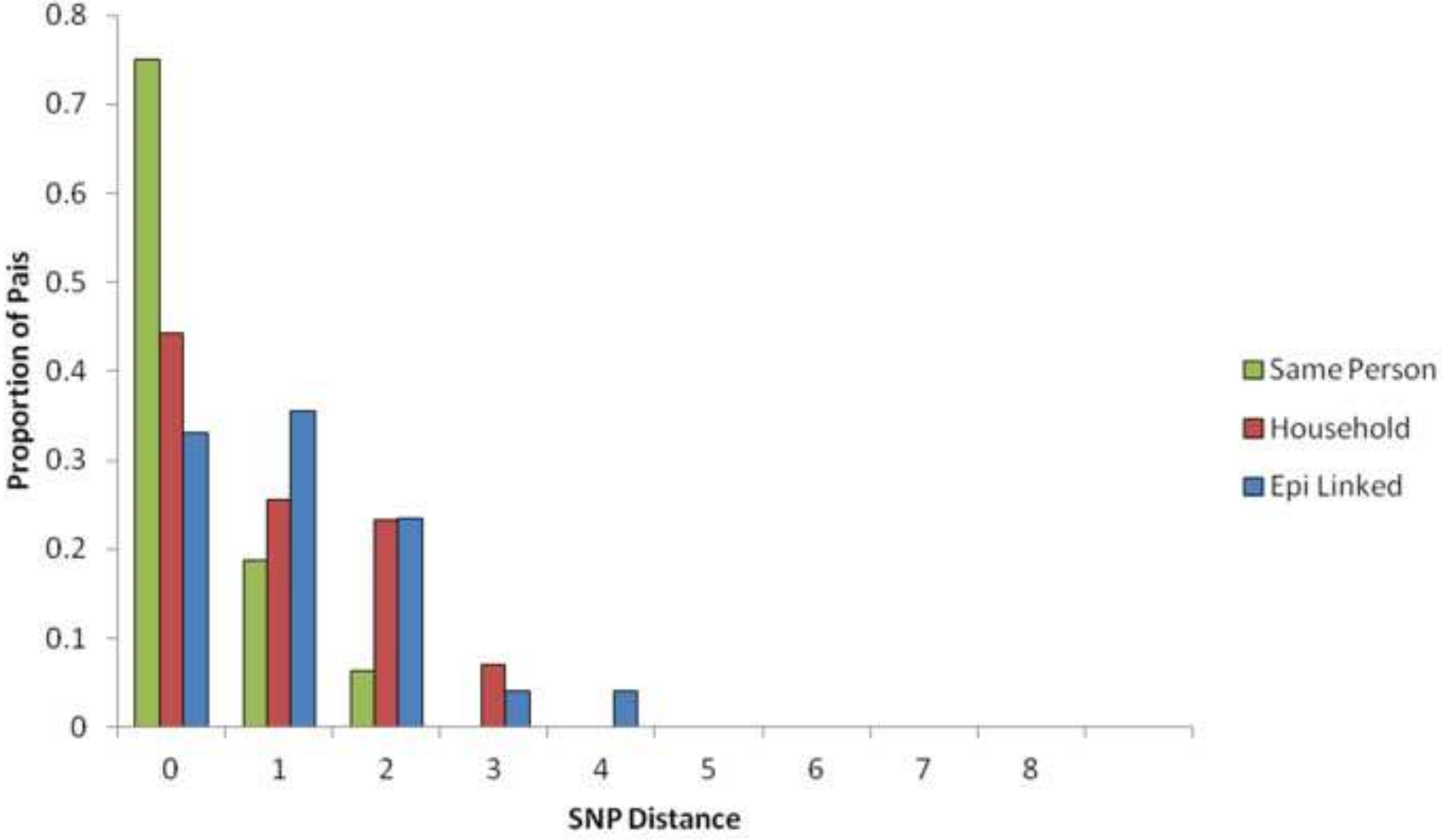


Figure 2

Click here to download high resolution image

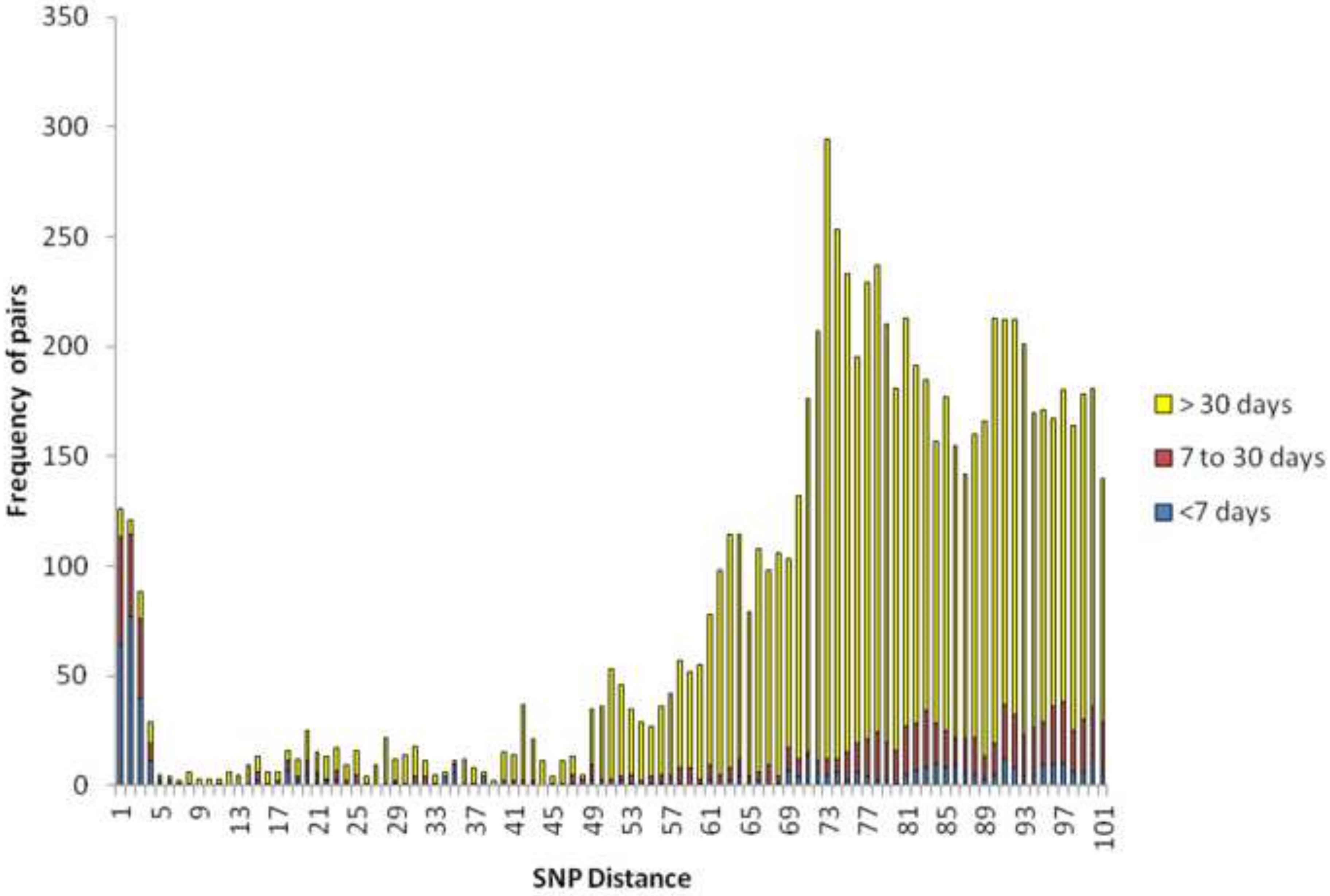


Figure 3a
Click here to download high resolution image

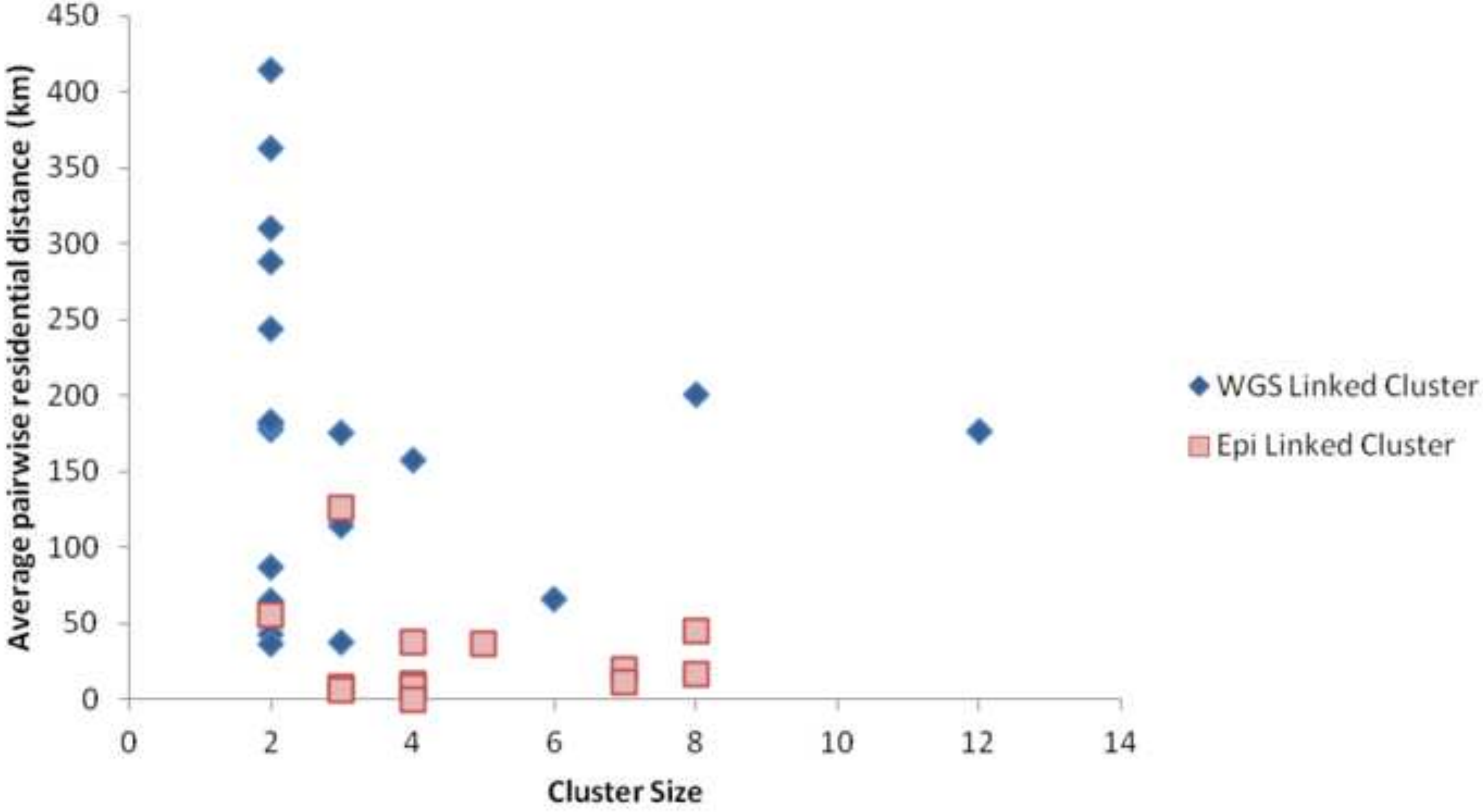


Figure 3b

Click here to download high resolution image

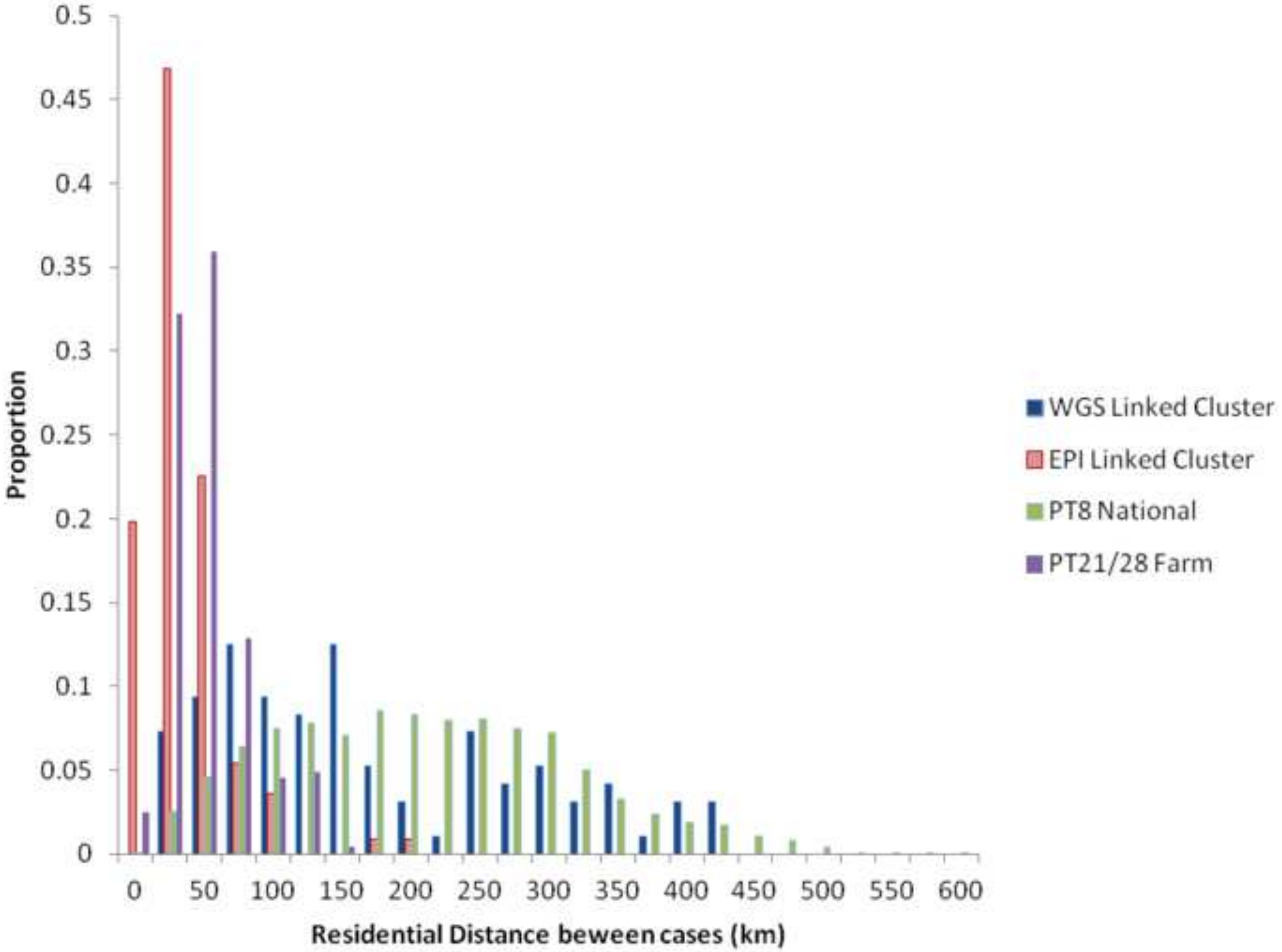


Figure 4

Click here to download high resolution image

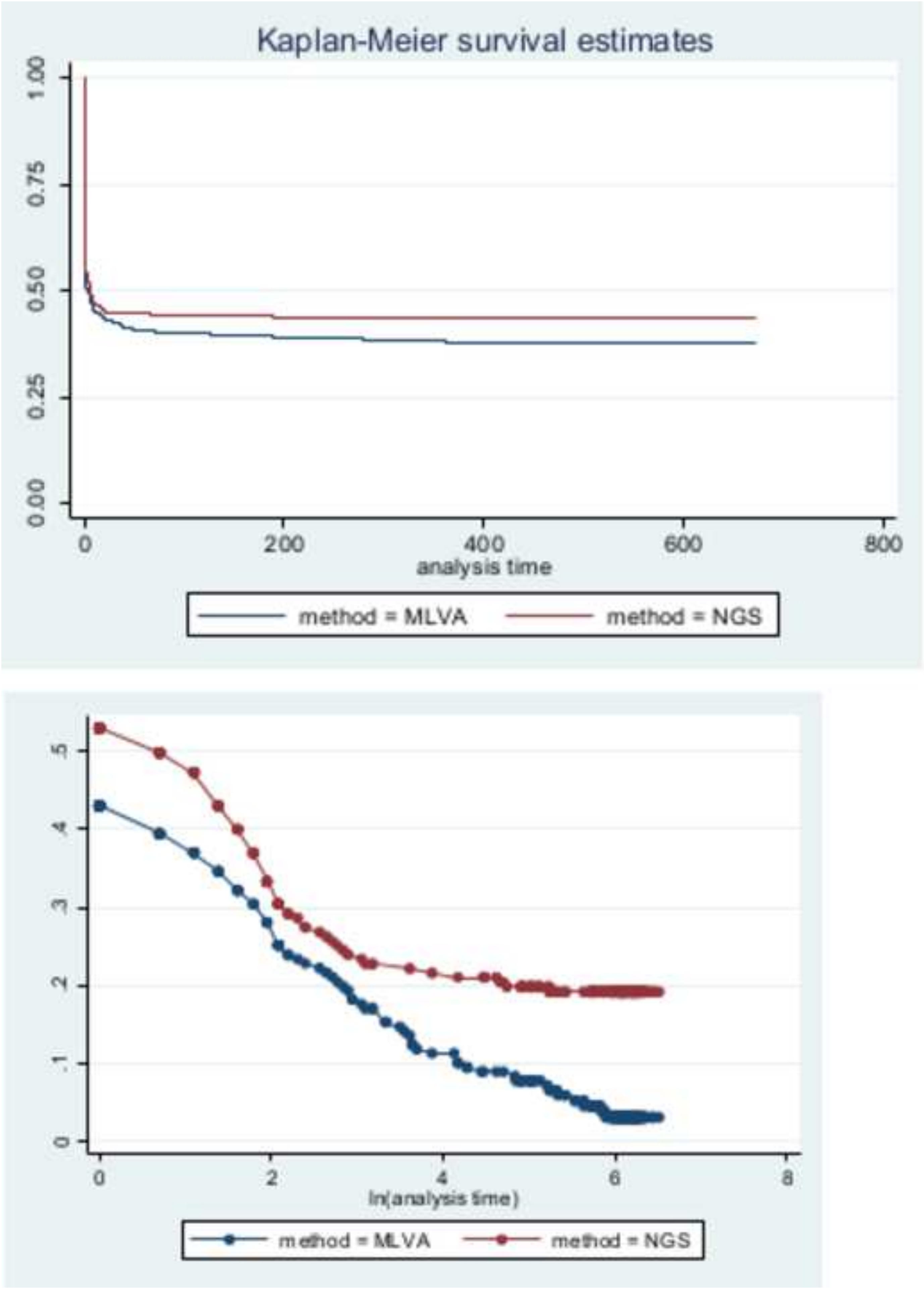
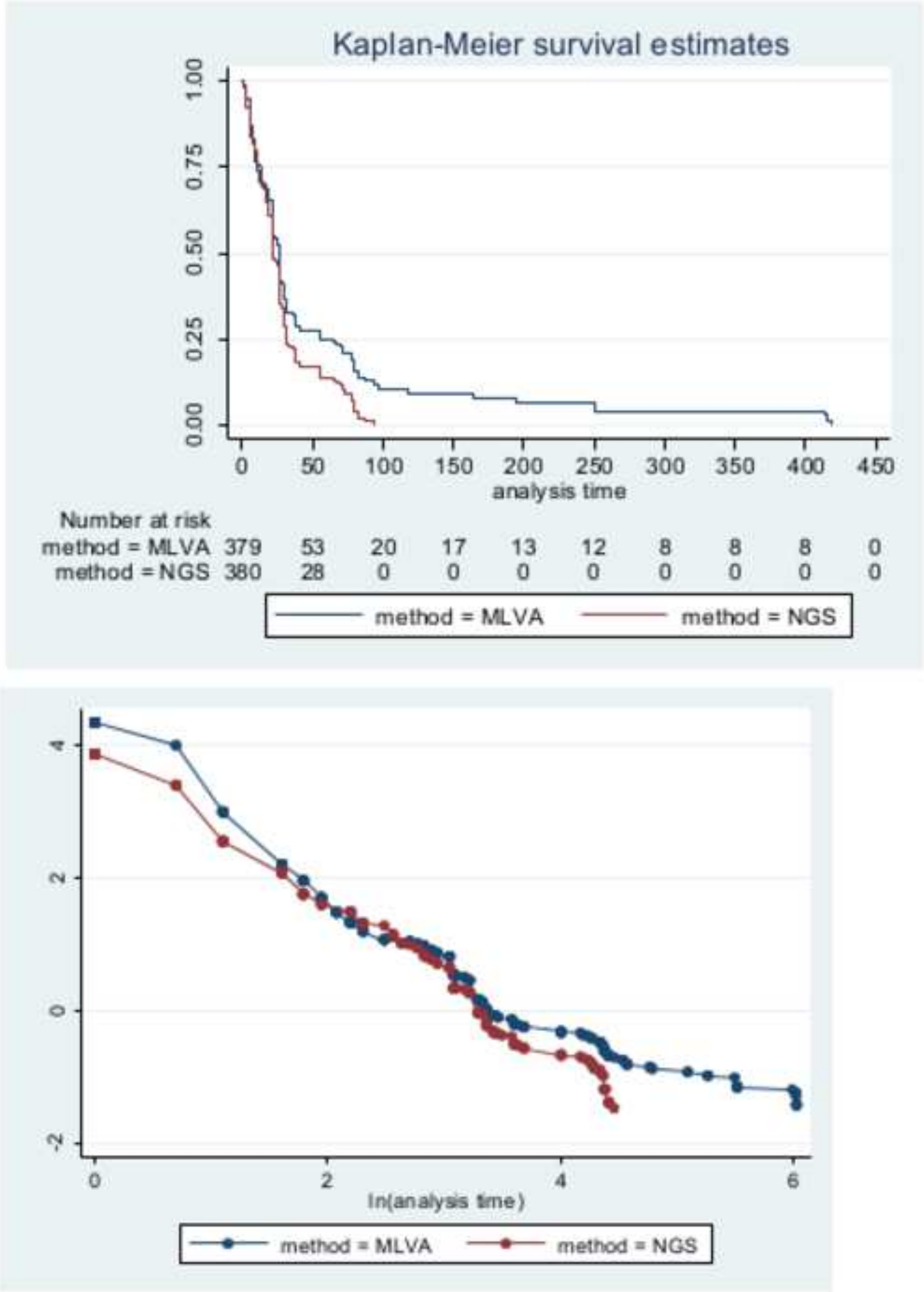


Figure 5

Click here to download high resolution image



Number at risk

method = MLVA	379	53	20	17	13	12	8	8	8	0
method = NGS	380	28	0	0	0	0	0	0	0	0

Figure 6

Click here to download high resolution image

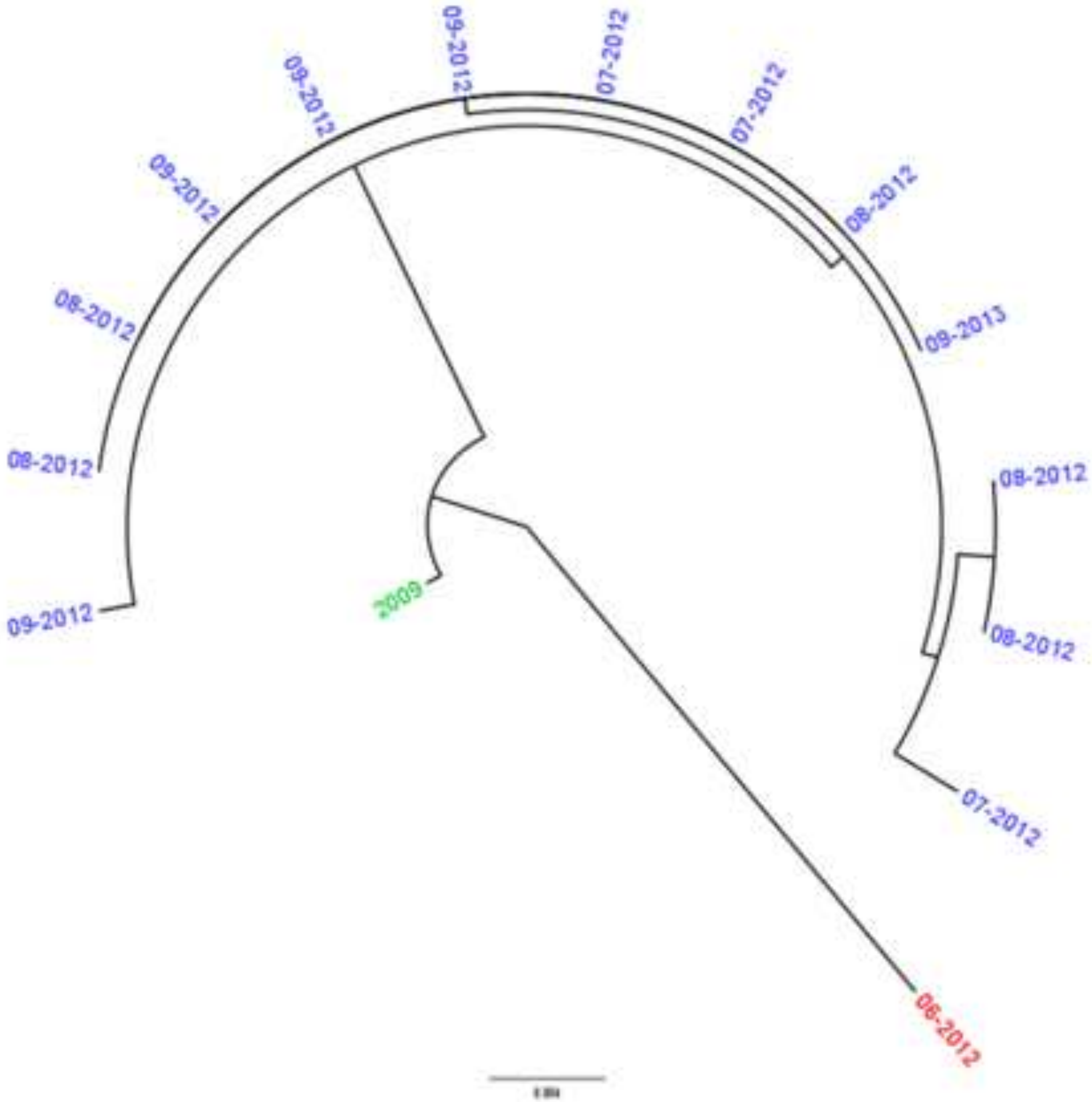
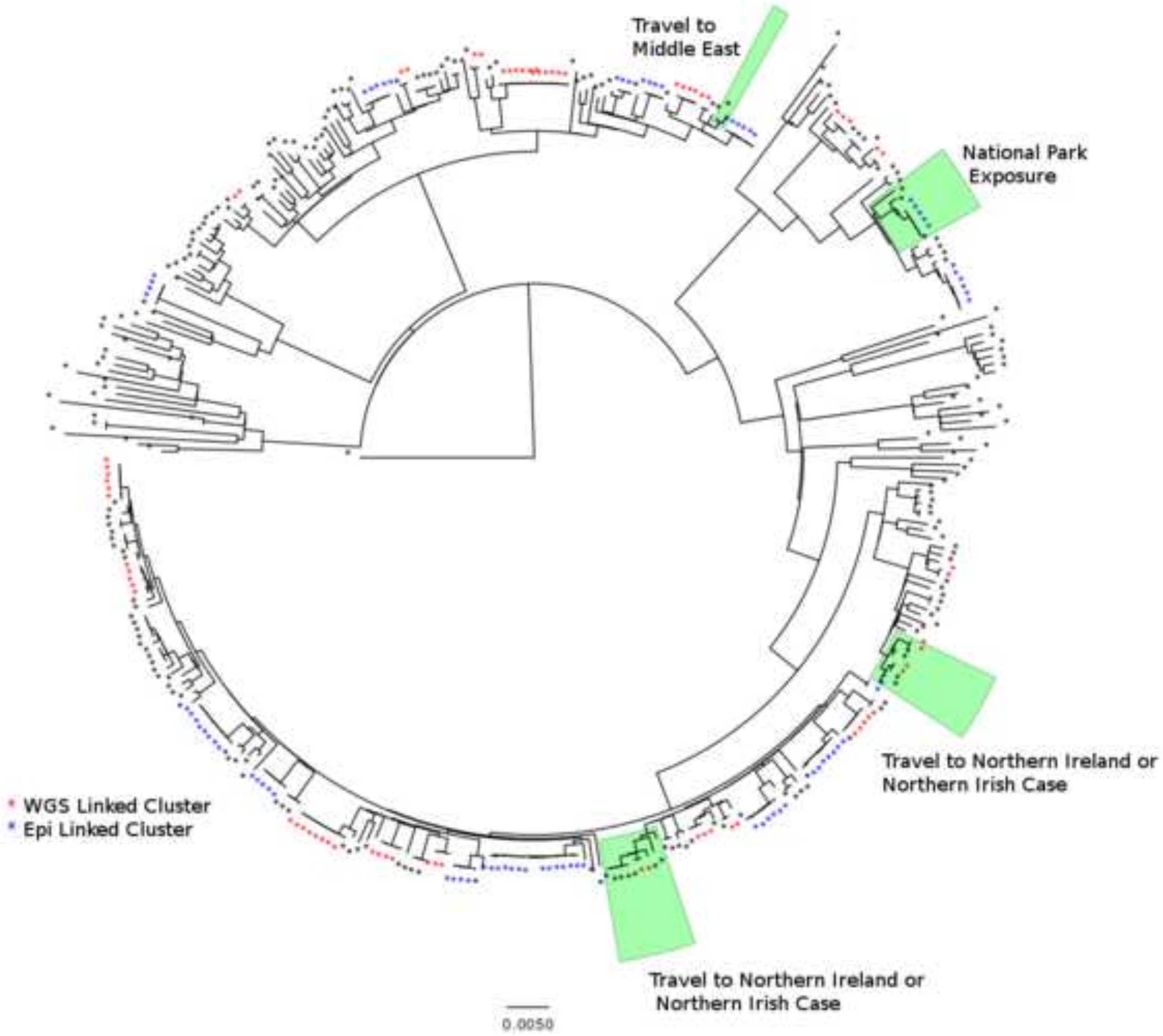


Figure 7

Click here to download high resolution image



Travel to
Middle East
National Park
Exposure
Travel to Northern Ireland or
Northern Irish Case
* WGS Linked Cluster
* Epi Linked Cluster
Travel to Northern Ireland or
Northern Irish Case
0.0050