Genomic Analysis of *Salmonella enterica* Serovar Typhimurium DT160 Associated with a 14-Year Outbreak, New Zealand, 1998–2012

Samuel J. Bloomfield, Jackie Benschop, Patrick J. Biggs,[1] Jonathan C. Marshall,[1] David T.S. Hayman,[1] Philip E. Carter,[1] Anne C. Midwinter, Alison E. Mather, Nigel P. French

During 1998–2012, an extended outbreak of *Salmonella enterica* serovar Typhimurium definitive type 160 (DT160) affected >3,000 humans and killed wild birds in New Zealand. However, the relationship between DT160 within these 2 host groups and the origin of the outbreak are unknown. Whole-genome sequencing was used to compare 109 *Salmonella* Typhimurium DT160 isolates from sources throughout New Zealand. We provide evidence that DT160 was introduced into New Zealand around 1997 and rapidly propagated throughout the country, becoming more genetically diverse over time. The genetic heterogeneity was evenly distributed across multiple predicted functional protein groups, and we found no evidence of host group differentiation between isolates collected from human, poultry, bovid, and wild bird sources, indicating ongoing transmission between these host groups. Our findings demonstrate how a comparative genomic approach can be used to gain insight into outbreaks, disease transmission, and the evolution of a multihost pathogen after a probable point-source introduction.

Nontyphoidal serovars of *Salmonella enterica* subsp. *enterica*, which cause salmonellosis, are responsible for an estimated 93.8 million illnesses and 155,000 deaths among humans worldwide each year (*1*). In New Zealand, these serovars are the second largest cause of bacterial gastroenteritis, annually causing 21 cases per 100,000 population (*2*). Nontyphoidal *Salmonella* spp. strains vary in host specificity and are usually transmitted to humans via direct contact or consumption of foods originating from animals (*3*,*4*). In New Zealand, salmonellosis incidence among humans peaks in the warm summer months, probably in association with increased multiplication of *Salmonella* in animal and food sources and with increased participation in higher risk outdoor activities (e.g., activities that increase contact with wild-life) (*5*). Climate change is expected to increase summer temperatures, potentially increasing salmonellosis incidence in New Zealand (*6*).

During 1998–2012, an extended outbreak of *Salmonella* Typhimurium definitive type 160 (DT160) occurred in New Zealand (*7*). During the outbreak, DT160 was the predominant *Salmonella* spp. subtype isolated from human salmonellosis patients and sick wild birds. DT160 was also isolated from other animals and the environment, but it was not the main *Salmonella* subtype isolated from these sources (*8*–*10*). DT160 has been isolated from animals and environments worldwide (*11*,*12*) and is usually associated with moribund birds (*13*,*14*). However, before the 1998–2012 outbreak, DT160 had not been reported in New Zealand. In 2009, an outbreak of DT160 involving humans and wild birds was reported in Tasmania, Australia (*15*); however, as with the outbreak in New Zealand, the relationship between DT160 within the bird and human host groups of Tasmania was unknown. We used genomic epidemiologic approaches to characterize the origin, evolution, and transmission of *Salmonella* Typhimurium DT160 in New Zealand.

Methods

Whole-Genome Sequencing

After stratifying the *Salmonella* strain collection at the Enteric Reference Laboratory of the Institute of Environmental Science and Research Ltd. (Wallaceville, New Zealand) by age and host, we randomly selected 35 human, 25 wild bird, 25 poultry, and 24 bovine DT160 isolates from 1998–2012. We extracted genomic DNA from these isolates

Author affiliations: Massey University, Palmerston North, New Zealand (S.J. Bloomfield, J. Benschop, P.J. Biggs, J.C. Marshall, D.T.S. Hayman, A.C. Midwinter, N.P. French); Institute of Environmental Science and Research, Wellington, New Zealand (P.E. Carter); University of Cambridge, Cambridge, UK (A.E. Mather)

DOI: https://dx.doi.org/10.3201/eid2306.161934

[1]These authors contributed equally to this article.

using a QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany) (*16*). New Zealand Genomics Limited (NZGL) at Massey Genome Service, Massey University, Palmerston North, New Zealand, performed whole-genome sequencing of the extracts. NZGL also prepared a library for each isolate by using a TruSeq DNA PCR-Free Library Preparation Kit (Illumina, Scorsby, Victoria, Australia) and sequenced the libraries by using MiSeq (Illumina, San Diego, CA, USA) as 2 × 250 bp paired-end runs (≈120–150 genome coverage). After sequencing and standard barcode demultiplexing, NZGL used FASTQ-MCF (*17*) to perform quality control procedures to remove any PhiX control library reads and adaptor sequences. The raw reads for the 109 DT160 isolates are available in the European Nucleotide Archive (http://www.ebi.ac.uk/ena; accession no. PRJEB18077).

Genomic Assembly

Each isolate's genome was assembled de novo. We used an in-house Perl script to trim reads at an error probability of 0.01 and generate random subsets of paired reads from 750,000 to 1.2 million paired reads in increments of 150,000, varying the average coverage. We assembled each of the random sets by using the de novo assembler Velvet version 1.1 (*18*) at a variety of k-mers (from 55 to 245) in increments of 10. De novo assembly resulted in multiple genome assemblies for each isolate. We ranked the metrics for each of 4 parameters (longest genome length, fewest number of contigs, largest N_{50} value, and longest contig length) in numeric order and calculated an overall equally summed ranking score for each assembly. We used the assemblies with the lowest total rank for further analyses. We used QUAST (*19*), a quality assessment tool for evaluating and comparing genome assemblies, to analyze the DT160 de novo assemblies and determine their GC content (i.e., the percentage of a DNA sequence made up of guanine and cytosine bases).

Single-Nucleotide Polymorphism Identification

We used Snippy version 2.6 (https://github.com/tseemann/snippy) and kSNP version 3.0 (*20*) to identify core single-nucleotide polymorphisms (SNPs). Snippy is a pipeline that uses the Burrows-Wheelers Aligner (*21*) and SAMtools version 1.3.1 (*22*) to align reads from different isolates to a sequence and uses FreeBayes (*23*) to identify variants among the alignments. We used kSNP to analyze de novo assembled genomes, along with the reference genome, *S. enterica* serovar Typhimurium 14028S (GenBank accession no. NC_016856). We used an in-house Python script to determine the read coverage of all the SNPs identified via kSNP. We used Snippy to align reads from each isolate to the reference genome (GenBank accession no. NC_016856) before identifying SNPs. SNPs were accepted if they had a >10 read depth and a >90% consensus for each isolate. The position of the SNP on the reference genome was used to determine if both methods identified the SNP or if they were unique to the method (online Technical Appendix, https://wwwnc.cdc.gov/EID/article/23/6/16-1934-Techapp1.pdf). This method identified 793 core SNPs shared by the 109 New Zealand DT160 isolates.

Global DT160 Strains

Using the genomic assembly and SNP identification methods as we described, we compared 2 DT160 strains from the United Kingdom with the 109 DT160 isolates from New Zealand: 1,521 core SNPs were identified. We downloaded the UK strains, which were previously published by Petrovska et al. (*24*), from the European Nucleotide Archive (accession nos. ERS015626 and ERS015627).

Phylogenetic Inference and Distances

We used RAxML version 8.2.4 (*25*) to construct a maximum-likelihood tree based on the 793 core SNPs of the 109 DT160 isolates; we used EvolView version 2 (*26*) to visualize and edit the tree. We used SplitsTree (*27*) to form a NeighborNet tree of the 109 New Zealand DT160 isolates based on the 793 core SNPs that they share and to compare the New Zealand and UK isolates based on the 1,521 core SNPs that they share. We used MEGA6 (*28*) and the maximum composite likelihood model (29) to predict the pairwise distance between the 109 New Zealand DT160 isolates, based on the 793 core SNPs they share, and the 109 New Zealand and 2 UK isolates, based on the 1,521 core SNPs that they share.

Phylogenetic Analysis

We used an in-house Perl script to split the 793 codons into 5 groups: those associated with the first, second, or third codon; those contained in overlapping coding regions; and those found in intergenic regions. We also used the in-house Perl script to determine whether the SNPs were synonymous or nonsynonymous. We then exported the partitioned SNPs into BEAUti to create an XML file for BEAST 1.8.3 (*30*).

To allow for variation in base substitution among codon positions, we used separate Hasegawa Kishino Yano models to estimate the 5 SNP groups (*31*); to allow for and estimate changes in the effective population size, we used the Gaussian Markov random field Bayesian skyride model (*32*); to allow for variation in mutation rates among lineages, we used an uncorrelated relaxed molecular clock (*33*), which was calibrated by the tip dates. We ran the XML file in BEAST for 40 million steps a total of 3 times with different starting seeds before using LogCombiner (http://beast.bio.ed.ac.uk/LogCombiner) to combine the runs with a 10% burn-in. To visualize the results and the relative change in effective population size, we used Tracer version 1.6 (*34*).

To determine the mutation rate for the DT160 genome, we multiplied the mutation rate estimated by BEAST by the number of analyzed core SNPs (793 bp) and then divided the product by the mean genome size of the analyzed isolates (4,884,485 bp). We used the discrete phylogeographic model (*35*) to predict ancestral migrations between host groups over the course of the outbreak.

Protein Coding Gene Analysis

We used Prokka (*36*) to annotate de novo assembled genomes, and we used Roary (*37*) to cluster proteins and identify those that were found only in a subset of isolates and those that differed in length between the isolates. We used ClustalW version 2.1 (*38*) to align amino acid sequences, and we used an in-house Perl script to determine if these alignments contained mismatches. The nucleotide sequence of all proteins that differed were extracted from the assembled genomes, along with 500-bp flanks on either side of the sequence, by using an in-house Perl script. We could not obtain 500-bp flanks for some genes because they were located at the end of contigs. For those genes, the flank was cut short, but their length was annotated. We extracted flanks to help with read alignment. This extraction left a pool of nucleotide sequences from each isolate, for every protein that potentially differed in sequence. For each protein, we extracted all nucleotide variants from the pool by using an in-house Perl script. We used SRST2 version 2, a read mapping–based tool (*39*), to align reads from each isolate to the sequence variants, and we used SAMtools version 1.3.1 (*22*) to form a consensus sequence from the aligned reads. We set the consensus cutoff at a read depth of ≥8 and a consensus of ≥80%. The flanks were removed from the consensus sequences, and the sequence variants were translated into amino acid sequences by using an in-house Perl script. We identified protein differences by comparing the amino acid sequences from each isolate and combined the differences with the nonsynonymous SNPs identified by SNP analysis. The position of nonsynonymous SNPs within proteins was used to prevent repeats.

We used the Clusters of Orthologous Groups of proteins (COGs) database (*40*) to predict protein functions. For each functional group, we calculated the proportion of proteins that differed in sequence, and we used a Fisher exact test, computed via Monte Carlo Markov Chains of $\approx 10^9$ iterations, to determine if there were any differences between these proportions.

We used an in-house Perl script to form a presence–absence matrix of all the protein differences. We used PrimerE version 6 (*41*) to predict the Euclidian distance between the isolates based on the presence–absence matrix. The centroid is the arithmetic mean for a group of data points in an n-dimensional space. To assess differences in centroids among isolates collected from different sources or time periods, we applied PERMANOVA (http://www.primer-e.com/permanova.htm). To assess differences in dispersions between different groups, we computed dispersions (z-values) by using PermDisp (*42*) and then modeled them using a regression model with date of collection and source as the explanatory variables.

Scripts

The in-house scripts used for genomic analyses in this study were specifically designed for this dataset. The scripts are available from GitHub (https://github.com/samuelbloomfield/Scripts-for-genomic-analyses).

Results

During the 1998–2012 human outbreak of *Salmonella* Typhimurium DT160 in New Zealand, disease incidence displayed a typical epidemic curve: prevalence increased from 1999 to 2000, before peaking at 791 cases in 2001, and then slowly decreased from 2002 through 2012 (Figure 1). At the same time, numerous isolates were reported from nonhuman hosts (wild birds, poultry, bovids), and disease incidence among these host groups displayed epidemic curves similar to those for humans (online Technical Appendix).

Genomic DT160 Comparison

The genomes we assembled were 4.8–4.9 Mb in length and had a GC content of 52.11%–52.16% (reference value for *S. enterica* 50%–53%) (*43*). We identified 793 core SNPs shared by the 109 DT160 isolates from New Zealand.

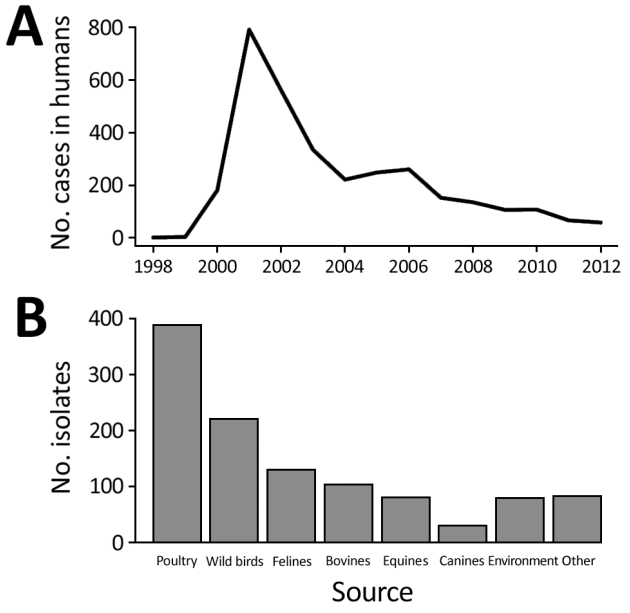


Figure 1. Number of *Salmonella enterica* serovar Typhimurium DT160 cases and isolates reported during an outbreak in New Zealand, 1998–2012. A) Cases in humans (*8*,*9*). B) Isolates from nonhuman sources (*8*,*10*).

DT160 Introduction Date

Ancestral date reconstruction analysis predicted that the 109 New Zealand DT160 isolates shared a date of common ancestor in approximately August 1997 (95% highest posterior density interval June 1996–August 1998). Comparative analysis indicated that the 2 DT160 isolates collected from the United Kingdom were genetically distinct from the 109 New Zealand DT160 isolates (online Technical Appendix). The average pairwise SNP distance between the 2 UK DT160 isolates and the New Zealand isolates was 0.0287, compared with an average pairwise distance of 0.0151 between New Zealand isolates.

In New Zealand, DT160 was first reported in Christchurch in 1998 from a human with salmonellosis (*44*) (an isolate from this case was included as part of this study). The New Zealand DT160 isolates we analyzed were estimated to share a common ancestor 0–2 years before this case and were distinct from the UK isolates analyzed, suggesting that DT160 was probably introduced into New Zealand as a single incursion within this time period. However, worldwide comparative studies are required to track DT160 migration and validate this hypothesis.

DT160 Evolution

Our phylogenetic analysis also predicted that the 109 DT160 isolates mutated at a rate of 3.3–4.3 × 10^{-7} substitutions/site/year (95% highest posterior density interval) and that the effective population size for DT160 increased from 1998 to 2003 (Figure 2). Over the course of the outbreak, DT160 also increased in genetic diversity (Figure 3).

The mutation rate estimated for the DT160 outbreak is similar to rates reported by Mather et al. (*45*) for an outbreak of *Salmonella* Typhimurium DT104 in Scotland during 1990–2012 and by Okoro et al. (*46*) for invasive *Salmonella* Typhimurium strains in sub-Saharan Africa. The similarity of these mutation rates suggests consistency between outbreaks caused by *S. enterica* serovar Typhimurium and has implications for modeling the evolution of future outbreaks caused by this serovar.

In bacteriology, the effective population size is the number of bacteria that contribute to the next generation. The increase in the DT160 effective population size during 1998–2003 coincided with an increased prevalence of DT160 among human and nonhuman hosts during this time. However, the subsequent levelling-off of the effective DT160 population size is probably an artifact because we calculated the effective population size from the timing of coalescent events for randomly sampled bacteria (*32*), and as the outbreak proceeded, fewer coalescent points were available for estimation.

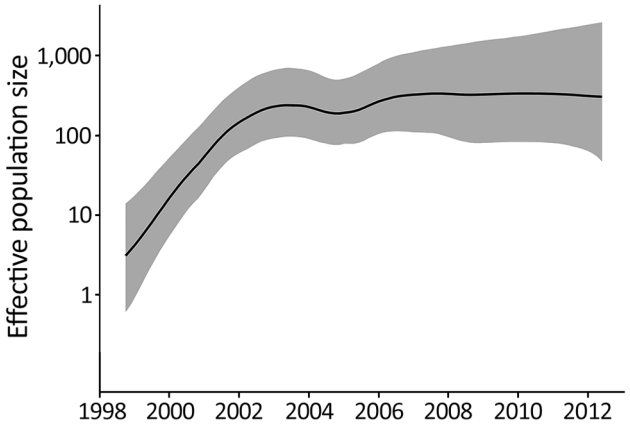


Figure 2. Relative effective population size (log scale) of *Salmonella enterica* serovar Typhimurium DT160 during an outbreak in New Zealand, 1998–2012. Population parameters were estimated using the Gaussian Markov random field Bayesian skyride model. The black line represents the median effective population size estimate; gray shading represents the 95% highest posterior density interval.

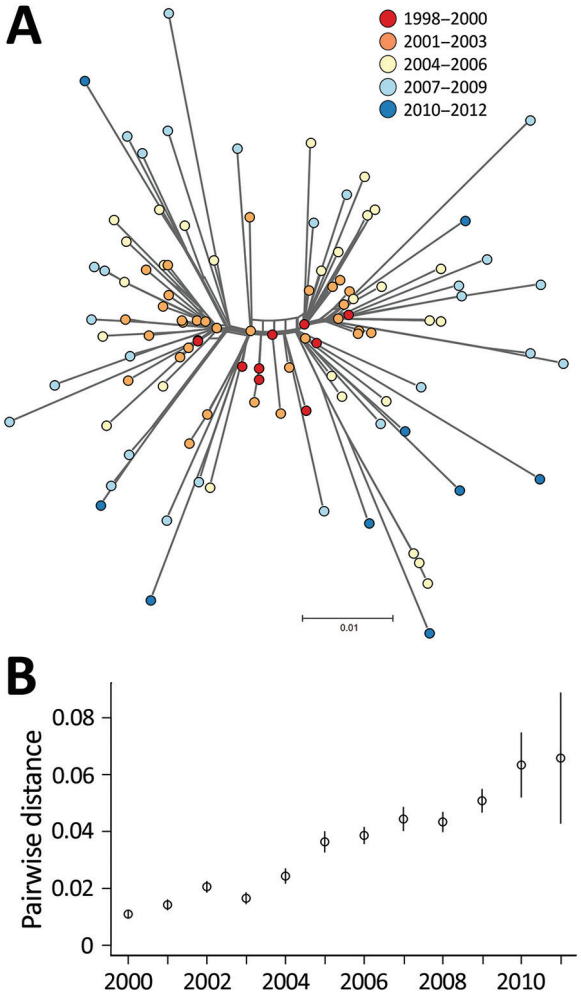


Figure 3. A) NeighborNet tree of 109 *Salmonella enterica* serovar Typhimurium DT160 isolates collected during an outbreak in New Zealand, 1998–2012. The tree was based on 793 core single-nucleotide polymorphisms. Colors indicate date of isolate collection. The scale bar represents the number of nucleotide substitutions per site. B) Scatterplot of the mean pairwise distance of 106 DT160 isolates from 2000–2011. Error bars represent 95% CIs.

Overall, our phylogenetic analyses suggest that the DT160 population increased dramatically in the first few years following introduction. As the DT160 population increased, it acquired multiple SNPs, resulting in a progressive increase in diversity over time.

DT160 Sources

PFGE (pulsed-field gel electrophoresis) was previously used to compare New Zealand DT160 isolates from humans, poultry, and wild birds (S. Omar, master's thesis, 2011; http://mro.massey.ac.nz/handle/10179/2681?show=full); however, PFGE could not distinguish DT160 from the separate sources. In our study, we were able to use whole-genome sequencing to distinguish DT160 at the isolate level. However, we did not find any distinct DT160 clades associated with any one source (Figure 4).

Identifying the source of a salmonellosis outbreak can be difficult because multiple potential sources must

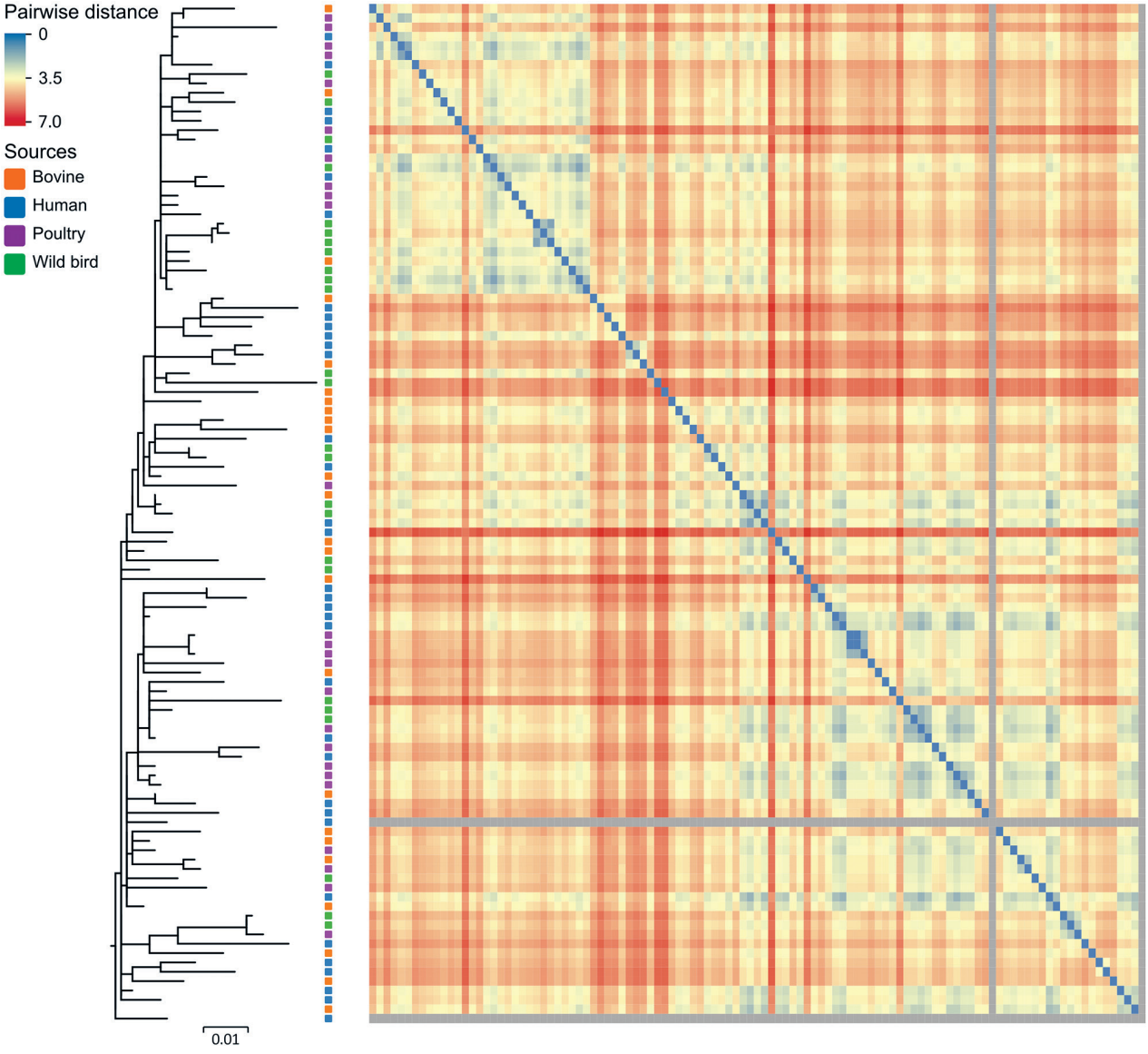


Figure 4. Maximum-likelihood tree of 109 *Salmonella enterica* serovar Typhimurium DT160 isolates collected during an outbreak in New Zealand, 1998–2012. The tree was based on 793 core single-nucleotide polymorphisms. Colored squares to the right of the branches indicate the source of isolates. The scale bar represents number of nucleotide substitutions per site. The heat map represents the Euclidean pairwise distance between isolates (based on the presence of 684 protein differences). Isolates that shared a small number of protein differences contained small Euclidean distances and are closer to blue in color on the heat map; isolates that shared a large number of protein differences contained large Euclidean distances and are closer to red in color. The gray squares represent the 2 outliers missing a large number of genes. The diagonal array of blue squares represents the pairwise distance for the same isolates.

be considered (*47*). Probable sources of *Salmonella* can be identified by comparing isolates from infected humans with those from other human, nonhuman, and environmental sources (*48*). We did not find distinct DT160 clades associated with any 1 source, suggesting that after its introduction into New Zealand, DT160 was transmitted between multiple hosts, resulting in large epidemics among humans and wild birds. Our results also suggest that humans obtained DT160 from multiple sources over the course of the outbreak. This finding is consistent with that in a case-control study performed by Thornley et al. (*44*), which found that human DT160 cases were associated with multiple risk factors involving different sources: handling dead wild birds, contact with persons with diarrhea, and consumption of fast food.

Ancestral Migration between Hosts

We used the discrete phylogeographic model to predict ancestral migration of DT160 between the animal and human host groups, similar to Mather et al. (*45*). However, we were unable to detect a signal that could not be attributed to different sampling fractions in the host groups (online Technical Appendix). Therefore, an alternate method, larger sample size, or both are required to predict these ancestral migrations.

Protein and Gene Analysis

Protein annotation identified 5,096 coding DNA sequences, of which 4,983 (98%) were found in all of the isolates, 108 (2%) were found in 95%–99% of isolates, and 3 (<1%) were found in 1%–5% of the isolates. Protein coding gene analysis also identified 477 nonsynonymous SNPs, of which 27 were nonsense mutations and 96 were INDELs (insertions/deletions). The nonsense SNPs and INDELs were responsible for 123 proteins that differed in length. Overall, we identified 684 differences in 604 protein sequences among the 109 DT160 isolates. We excluded 2 isolates from protein coding gene analysis because they were missing a large number of proteins (online Technical Appendix).

By using PERMANOVA, we found that centroids based on the 684 protein differences were indistinguishable among groups of DT160 isolates collected from different sources and time periods (online Technical Appendix). PERMANOVA's inability to distinguish centroids appears to be due to the fact that DT160 isolates radiated out from a point source. The z-value is the distance from an isolate to the centroid of a group of isolates; we calculated the z-value for 107 DT160 isolates on the basis of 684 protein differences. Our regression modeling results showed that the z-value was associated with the date, but not source, of collection (Figure 5).

The 684 protein differences shared by the DT160 isolates were associated with a large number of COG functional groups. The proportion of proteins that contained sequence differences differed between functional groups ($p = 0.00002$). The proportions varied from 0.06 to 0.18, although most were between 0.09 and 0.13 (online Technical Appendix). In addition, our data were insufficient to model the effects of source or date of collection on the number of protein differences associated with each group (online Technical Appendix).

Bacteria often adapt to new environments by altering (changing or losing) genes that are not essential for colonizing that environment (*49*). Gene loss can result in an increase in bacterial fitness, as fewer genes and processes need to be maintained within the bacteria (*50*). We

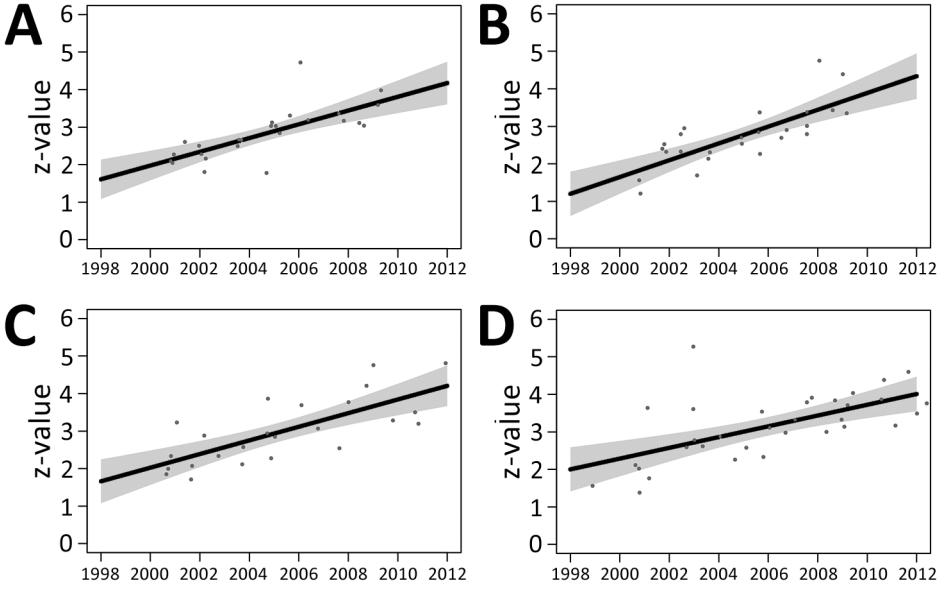


Figure 5. Scatter plots of year of collection versus z-values for 107 Salmonella enterica serovar Typhimurium DT160 isolates collected during an outbreak in New Zealand, 1998–2012. Of the 107 isolates, 25 were from poultry (A), 25 from wild birds (B), 24 from bovids (C), and 33 from humans (D). Black lines represent the regression equation; gray shading represents SE for this equation. Date of collection was significantly associated with z-values in this model ($p<2-16$). There was insufficient evidence to suggest that source was associated with z-values ($p = 0.558$), and the interaction between source and date of collection was not significant ($p = 0.458$).

identified multiple protein changes among the DT160 isolates, and these changes occurred in multiple COG functional groups as the epidemic progressed. However, we found no evidence of host group differentiation, suggesting that most of the evolution was due to random genetic drift rather than adaptive evolution.

Discussion

Using genomic analysis, we described the evolution and emergence of *Salmonella* Typhimurium DT160 within New Zealand. Our results suggest that DT160 was introduced into New Zealand on a single occasion from 1996 through 1998, before propagating throughout the country and becoming more genetically diverse over time. In addition, we found that DT160 isolates collected from human, poultry, bovine and wild bird sources were highly similar, indicating a large number of transmission episodes between these host groups.

Acknowledgments

We thank the staff at the Massey Genome Service (part of New Zealand Genomics Ltd), Massey University, Palmerston North, New Zealand, for their help and advice with the genome sequencing for this study; Tim G. Vaughan for his help with phylogenetic analyses; and the Enteric Reference Laboratory, Environmental Science and Research, Wallaceville, New Zealand for supplying *Salmonella* isolates.

This study was funded by the Allan Wilson Centre for Molecular Ecology and Evolution, Massey University; the Biotechnology and Biological Sciences Research Council (BB/M014088/1 to A.E.M); and the Royal Society of New Zealand Marsden Fund (MAU1503 to D.T.S.H.).

Mr. Bloomfield is a PhD student at Massey University, Palmerston North, New Zealand. His research interests include infectious diseases and genomics, particularly how infectious diseases are transmitted and evolve over the course of outbreaks.

References

1. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, et al.; International Collaboration on Enteric Disease 'Burden of Illness' Studies. The global burden of nontyphoidal *Salmonella* gastroenteritis. Clin Infect Dis. 2010;50:882–9. http://dx.doi.org/10.1086/650733
2. Institute of Environmental Science and Research Ltd. (ESR). Notifiable diseases. New Zealand public health observatory. 2014 [cited 2016 Feb 2]. http://www.nzpho.org.nz/NotifiableDisease.aspx
3. King N, Lake R, Campbell D. Source attribution of nontyphoid salmonellosis in New Zealand using outbreak surveillance data. J Food Prot. 2011;74:438–45. http://dx.doi.org/10.4315/0362-028X.JFP-10-323
4. Eswarappa SM, Janice J, Nagarajan AG, Balasundaram SV, Karnam G, Dixit NM, et al. Differentially evolved genes of *Salmonella* pathogenicity islands: insights into the mechanism of host specificity in *Salmonella.* PLoS One. 2008;3:e3829. http://dx.doi.org/10.1371/journal.pone.0003829
5. Lal A, Baker MG, French NP, Dufour M, Hales S. The epidemiology of human salmonellosis in New Zealand, 1997–2008. Epidemiol Infect. 2012;140:1685–94. http://dx.doi.org/10.1017/S0950268811002470
6. Lal A, Ikeda T, French N, Baker MG, Hales S. Climate variability, weather and enteric disease incidence in New Zealand: time series analysis. PLoS One. 2013;8:e83484. http://dx.doi.org/10.1371/journal.pone.0083484
7. Alley MR, Connolly JH, Fenwick SG, Mackereth GF, Leyland MJ, Rogers LE, et al. An epidemic of salmonellosis caused by *Salmonella* Typhimurium DT160 in wild birds and humans in New Zealand. N Z Vet J. 2002;50:170–6. http://dx.doi.org/10.1080/00480169.2002.36306
8. Institute of Environmental Science and Research Ltd (ESR). ESR LabLink. Quarterly surveillance summaries for New Zealand, March 2000–March 2003 [cited 2016 Nov 25]. https://surv.esr.cri.nz/PDF_surveillance/Lablink/
9. Institute of Environmental Science and Research Ltd (ESR). Public Health Surveillance; Information for New Zealand public health action. 2003–2012 human *Salmonella* isolates [cited 2016 Nov 25]. https://surv.esr.cri.nz/enteric_reference/human_salmonella.php
10. Institute of Environmental Science and Research Ltd (ESR). Public Health Surveillance; Information for New Zealand public health action. Non-human *Salmonella* isolates, 2003–2012 [cited 2016 Nov 25]. https://surv.esr.cri.nz/enteric_reference/nonhuman_salmonella.php
11. Penfold JB, Amery HC, Peet PJ. Gastroenteritis associated with wild birds in a hospital kitchen. Br Med J. 1979;2:802.
12. Tizard IR, Fish NA, Harmeson J. Free flying sparrows as carriers of salmonellosis. Can Vet J. 1979;20:143–4.
13. Lawson B, Howard T, Kirkwood JK, Macgregor SK, Perkins M, Robinson RA, et al. Epidemiology of salmonellosis in garden birds in England and Wales, 1993 to 2003. EcoHealth. 2010;7:294–306. http://dx.doi.org/10.1007/s10393-010-0349-3
14. Piccirillo A, Mazzariol S, Caliari D, Menandro ML. *Salmonella* Typhimurium phage type DT160 infection in two Moluccan cockatoos (*Cacatua moluccensis*): clinical presentation and pathology. Avian Dis. 2010;54:131–5. http://dx.doi.org/10.1637/8969-062509-Case.1
15. Grillo T, Post L. *Salmonella* Typhimurium DT160 outbreak in Tasmania. Animal Health Surveillance Quarterly Reports. 2010;14:8–8. http://www.sciquest.org.nz/node/72986
16. QIAGEN. QIAamp DNA mini and blood mini handbook. Third edition. 2012 [cited 2015 Feb 12]. https://moodle.ufsc.br/pluginfile.php/1379318/mod_resource/content/0/QIAamp_DNA_Mini_Blood.pdf
17. Aronesty E. Comparison of sequencing utility programs. The Open Bioinformatics Journal. 2013;7:1–8. http://dx.doi.org/10.2174/1875036201307010001
18 Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. http://dx.doi.org/10.1101/gr.074492.107
19. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5. http://dx.doi.org/10.1093/bioinformatics/btt086
20. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015;31:2877–8. http://dx.doi.org/10.1093/bioinformatics/btv271
21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60. http://dx.doi.org/10.1093/bioinformatics/btp324
22. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27:2987–93. http://dx.doi.org/10.1093/bioinformatics/btr509

23. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012 [cited 2017 Feb 2]. https://arxiv.org/abs/1207.3907
24. Petrovska L, Mather AE, AbuOun M, Branchu P, Harris SR, Connor T, et al. Microevolution of monophasic *Salmonella* Typhimurium during epidemic, United Kingdom, 2005–2010. Emerg Infect Dis. 2016;22:617–24. http://dx.doi.org/10.3201/eid2204.150531
25. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3. http://dx.doi.org/10.1093/bioinformatics/btu033
26. He Z, Zhang H, Gao S, Lercher MJ, Chen WH, Hu S. Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res. 2016;44:W236-41. http://dx.doi.org/10.1093/nar/gkw370
27. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67. http://dx.doi.org/10.1093/molbev/msj030
28. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013;30:2725–9. http://dx.doi.org/10.1093/molbev/mst197
29. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A. 2004;101:11030–5. http://dx.doi.org/10.1073/pnas.0404206101
30. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73. http://dx.doi.org/10.1093/molbev/mss075
31. Hasegawa M, Kishino H, Yano T. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74. http://dx.doi.org/10.1007/BF02101694
32. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008;25:1459–71. http://dx.doi.org/10.1093/molbev/msn090
33. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4:e88. http://dx.doi.org/10.1371/journal.pbio.0040088
34. Rambaut A, Suchard MA, Xie D, Drummond AJ. Beast. Tracer 1.6 2014 [cited 2016 Sep 27]. http://beast.bio.ed.ac.uk/Tracer
35. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLOS Comput Biol. 2009;5:e1000520. http://dx.doi.org/10.1371/journal.pcbi.1000520
36. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9. http://dx.doi.org/10.1093/bioinformatics/btu153
37. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3. http://dx.doi.org/10.1093/bioinformatics/btv421
38. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80. http://dx.doi.org/10.1093/nar/22.22.4673
39. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6:90. http://dx.doi.org/10.1186/s13073-014-0090-6
40. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28:33–6. http://dx.doi.org/10.1093/nar/28.1.33
41. Clark KR, Gorle RN. PRIMER v6: user manual/tutorial. Plymouth (UK): PRIMER-E; 2006. p. 296.
42. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 2006;62:245–53. http://dx.doi.org/10.1111/j.1541-0420.2005.00440.x
43. Popoff MY, Le Minor LE. Genus XXXIII. *Salmonella*. In: Brenner DJ, Staley JT, editors. Bergey's manual of systematic bacteriology. New York: Springer; 2005. p. 764–99.
44. Thornley CN, Simmons GC, Callaghan ML, Nicol CM, Baker MG, Gilmore KS, et al. First incursion of *Salmonella enterica* serotype Typhimurium DT160 into New Zealand. Emerg Infect Dis. 2003;9:493–5. http://dx.doi.org/10.3201/eid0904.020439
45. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, et al. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. Science. 2013;341:1514–7. http://dx.doi.org/10.1126/science.1240578
46. Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani MN, et al. Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. Nat Genet. 2012;44:1215–21. http://dx.doi.org/10.1038/ng.2423
47. Gieraltowski L, Julian E, Pringle J, Macdonald K, Quilliam D, Marsden-Haug N, et al. Nationwide outbreak of *Salmonella* Montevideo infections associated with contaminated imported black and red pepper: warehouse membership cards provide critical clues to identify the source. Epidemiol Infect. 2013;141:1244–52. http://dx.doi.org/10.1017/S0950268812001859
48. Byrne L, Fisher I, Peters T, Mather A, Thomson N, Rosner B, et al.; International Outbreak Control Team. A multi-country outbreak of *Salmonella* Newport gastroenteritis in Europe associated with watermelon from Brazil, confirmed by whole genome sequencing: October 2011 to January 2012. Euro Surveill. 2014;19: 20866. http://dx.doi.org/10.2807/1560-7917.ES2014.19.31.20866
49. Hottes AK, Freddolino PL, Khare A, Donnell ZN, Liu JC, Tavazoie S. Bacterial adaptation through loss of function. PLoS Genet. 2013;9:e1003617. http://dx.doi.org/10.1371/journal.pgen.1003617
50. Koskiniemi S, Sun S, Berg OG, Andersson DI. Selection-driven gene loss in bacteria. PLoS Genet. 2012;8:e1002787. http://dx.doi.org/10.1371/journal.pgen.1002787

Address for correspondence: Samuel J. Bloomfield, Massey University, Palmerston North, New Zealand; email: s.bloomfield@massey.ac.nz

Article DOI: https://dx.doi.org/10.3201/eid2306.161934

Genomic Analysis of *Salmonella enterica* Serovar Typhimurium DT160 Associated with a 14-Year Outbreak, New Zealand, 1998–2012

Technical Appendix

Sample Collection

From 1998–2012, *Salmonella enterica* serovar Typhimurium DT160 was isolated from humans and numerous animal and environmental sources in New Zealand. In this study, 35 human, 25 wild bird, 25 poultry and 24 bovine DT160 isolates were randomly selected from those isolates reported to the culture collection center at the Institute of Environmental Science and Research (ESR). The number of isolates reported in these host groups displayed similar epidemic curves, with an increase in prevalence from 1999–2000, before peaking in 2001 and slowly decreasing in prevalence from 2002–2012. (Technical Appendix Figure 1).

SNP Comparison

SNPs (single nucleotide polymorphisms) are single base pairs that differ between isolates. Two software programs were used to identify SNPs shared by the 109 DT160 isolates: Snippy (https://github.com/tseeman/snippy) and kSNP3 (*1*). Snippy was used to align reads from each isolate to a reference genome, in this case *S. enterica* serovar Typhimurium strain 14028s (NC_016856), and then to compare the alignment results and identify single base pairs that were found in all isolates but differed in sequence (core SNPs). kSNP was used to identify kmers of a fixed length that differed in one nucleotide between de novo-assembled genomes and NC_016856. kSNP identified 731 SNPs shared by the 109 DT160 isolates, while Snippy identified 771 SNPs (Technical Appendix Figure 2). 709 SNPs were identified by both methods, leaving 22 kSNP-unique and 62 Snippy-unique SNPs. The kSNP-unique SNPs mostly consisted of SNPs found on reads that did not align to the reference genome, while the Snippy-unique

SNPs mostly consisted of SNPs that were in close vicinity, unable to be picked up by kSNP as kmers of a fixed length would differ in more than one nucleotide. By using both methods a larger number of SNPs were identified than if a single method alone was used.

773 out of the 793 core SNPs shared by the 109 DT160 isolates were also located on the reference genome, NC_016856. The order of these SNPs on the reference genome identified several small clades associated with close clusters of SNPs (Technical Appendix Figure 3). However, most of the SNPs in these clusters were synonymous and unlikely to result from selection pressures. The order of these SNPs also identified the non-synonymous SNPs responsible for the formation of two distinct DT160 clades and the proteins they were located within: glycogen debranching enzyme (A), 2-dehydro-3-deoxyphosphooctonate aldolase (B), a YggT family protein (C), galactose-1-epimerase (D), uvrABC system protein B (E) and acrylyl-coA reductase (F). Many of these proteins are involved in carbohydrate metabolism, suggesting that the two DT160 clades may have distinct carbohydrate metabolism phenotypes.

Global DT160 Strains

Petrovska *et al*. (*2*) previously published the genomes of two DT160 isolates: ERS015626 that was isolated from a horse in 1998 and ERS015627 that was isolated from a bird in 1997. The raw reads from these isolates were downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena) and assembled de novo. kSNP and Snippy identified 1,521 core SNPs in total shared by these two isolates and the 109 New Zealand DT160 isolates analyzed. The average pairwise SNP distance between the two UK DT160 isolates and the New Zealand isolates was 0.0287, compared to an average pairwise distance between NZ isolates of 0.0151

The two DT160 isolates from the United Kingdom were genetically distinct from each other and from the 109 New Zealand DT160 isolates (Technical Appendix Figure 4). To our knowledge these were the only DT160 isolates published to date.

Protein Coding Gene Analysis

The 109 DT160 isolates shared 684 protein differences. Primer-E v6 (*3*) was used to predict the Euclidian distance matrix based on the presence of these protein differences.

Of the 684 proteins that differed in sequence, 546 (93%) contained a single protein difference (SNP, indel or presence), 53 (7%) contained two protein differences, and 5 (<1%) contained more than two (Technical Appendix Figure 5).

Two isolates were excluded from protein analyses as they lacked a large number of genes and were skewing the multi-dimensional scaling, functional plots and PermDisp calculations (Technical Appendix Figure 6). These outliers shared similar epidemiologic information: collected from human sources from 2004–2006. However, they were missing different sets of genes.

Multidimensional scaling helps visualize the amount of similarity or dissimilarity between data points. In multi-dimensional scaling, the centroid is the central point for a group of data points. PERMANOVA found that the centroids were indistinguishable between isolates collected from different sources or time periods (Table), as these isolates appeared to radiate out from a point source (Technical Appendix Figure 7).

The distance from the centroid to each isolate (z-value) is a measure of dispersion and equivalent to the accumulation of protein differences. The z-values were calculated using PermDisp (*4*) and were modeled using a regression model. The residuals for this model lacked normality (Technical Appendix Figure 8). To normalize the residuals, the z-values could have been transformed. However, with such a low p-value for the date of collection, this would not have changed the conclusions and would have made interpretation more difficult.

The 684 protein differences shared by the DT160 isolates were associated with a large number of functions. For each COG functional group, the proportion of proteins that differed in sequence was calculated (Technical Appendix Figure 9). Fisher exact test provided evidence that these proportions differed ($p = 0.0002$). However, there was little variation in the proportions (range: 0.07–0.17) and there were no outliers.

The mean proportion of proteins that differ in sequence for each functional group within each time period and source was calculated by dividing the proportion of proteins that differed in sequence among each source and time period in each functional group by the number of samples in each group (Technical Appendix Figures 10). Year of collection and source seemed to have a significant effect on the mean proportion of proteins that differ in sequence within each functional group: the proportion within each functional group tended to increase over time, and

certain functional groups (e.g., Extracellular structures (COG group W), Cell cycle control, cell division and chromosome partitioning (COG group D), Signal transduction mechanisms (COG group T), Lipid transport and metabolism (COG group I), and Cell motility (COG group N)) had higher proportions in the bovine and human host groups compared to the poultry and wild bird. However, the total number of protein differences within each functional group was smaller than the total number of samples (Technical Appendix Figure 11). Therefore, a regression model could not be used to model the effect of source and date of collection on the number of differences in each functional group, as a large number of isolates would have the same z-value.

Discrete Phylogeographic Model

The discrete phylogeographic model was designed to use phenotypic or molecular data to predict the ancestral migration of organisms from distinct geographies (*5*). However, the model has been applied to outbreaks to predict transmission between distinct host groups that share the same geography (*6*). Twenty-two datasets were formed from the 109 DT160 isolates and the 793 core SNPs they share, to determine if the discrete phylogeographic model was appropriate for investigating this outbreak. The real dataset consisted of the 109 isolates split into those from animal sources (n = 74) and those from human sources (n = 35) (real dataset). Ten datasets were formed by randomly assigning the 109 isolates as animal or human, while keeping the total number of animal and human isolates the same (datasets A-J). Eleven datasets were formed by randomly assigning one of the isolates as human, while assigning the rest as animal, before progressively assigning random isolates as human, until a range of data was formed with different numbers of human and animal isolates. Each dataset was exported into BEAUti to create an .xml file for BEAST 1.8.3 (*7*). For simplicity's sake, each dataset was given a separate Hasegawa Kishino Yano (HYK) substitution model (*8*) and strict molecular clock. The GMRF Bayesian skyride model (*9*) was used to allow for variation in the effective population size of each model and the discrete phylogeographic model (*5*) was used to predict the time spent in the animal and human host groups (Markov rewards) over the course of the outbreak, and the number of transmission between these host groups (Markov jumps). Each .xml file was run in BEAST for 10 million steps.

The discrete phylogeographic model predicted that DT160 spend most of the time in the animal host group, and that there was a larger amount of transmission from the animal to the human host group than the reciprocal. However, the same result was obtained when the isolates were randomly assigned as human or animal, but the sample proportions were kept the same (Technical Appendix Figures 12 and 13). In addition, the proportion of samples assigned as human had a significant effect on the Markov rewards and jumps (Technical Appendix Figures 14 and 15). This indicates that the results obtained from the discrete phylogeographic model are the result of an uneven sample size and not true migration events.

The proportion of samples that are human and Markov rewards share a step-like or sigmoid association (Technical Appendix Figure 14). This is due to the deep DT160 branches that are predominantly one source until the proportion of samples that are human meets a threshold (30%–40% of samples are human), where they suddenly all switch (Technical Appendix Figure 16). However, the relationship between the proportion of samples that are human and Markov jumps is more complex (Technical Appendix Figure 15). As the proportion of samples that are human increases, the number of human branches increases, but the ancestral branches remain animal, resulting in an increase in the number of animal-to-human Markov jumps. There are no human-to-animal Markov jumps up until the threshold, as there are no ancestral branches that are human. However, after the human proportion threshold is meet, the ancestral branches switch to human, resulting in no animal-to-human Markov jumps and a large number of human-to-animal Markov jumps that decrease as the human sample proportion increases and the number of animal tips decrease. If there were no deep branches or coalescent events, we would expect the correlation between the human proportion and Markov rewards to be more linear. In addition, we would expect there to be a positive linear relationship between the human proportion and the number of each Markov jump up to the threshold and a negative linear relationship afterwards.

References

1. Gardner SN, Slezak T, Hall BG. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics. 2015;31:2877–8. PubMed http://dx.doi.org/10.1093/bioinformatics/btv271

2. Petrovska L, Mather AE, AbuOun M, Branchu P, Harris SR, Connor T, et al. Microevolution of monophasic *Salmonella* Typhimurium during epidemic, United Kingdom, 2005–2010. Emerg Infect Dis. 2016;22:617–24. 10.3201/eid2204.150531 PubMed http://dx.doi.org/10.3201/eid2204.150531

3. Clark KR, Gorle RN. PRIMER v6: user manual/tutorial. Plymouth (UK): PRIMER-E; 2006. p. 296.

4. Anderson MJ. Distance-based tests for homogeneity of multivariate dispersions. Biometrics. 2006;62:245–53. PubMed http://dx.doi.org/10.1111/j.1541-0420.2005.00440.x

5. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLOS Comput Biol. 2009;5:e1000520. PubMed http://dx.doi.org/10.1371/journal.pcbi.1000520

6. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, et al. Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. Science. 2013;341:1514–7. PubMed http://dx.doi.org/10.1126/science.1240578

7. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29:1969–73. PubMed http://dx.doi.org/10.1093/molbev/mss075

8. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22:160–74. PubMed http://dx.doi.org/10.1007/BF02101694

9. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008;25:1459–71. PubMed http://dx.doi.org/10.1093/molbev/msn090

10. Institute of Environmental Science and Research Ltd (ESR). ESR LabLink. Quarterly surveillance summaries for New Zealand, March 2000–March 2003 [cited 2016 Nov 25]. https://surv.esr.cri.nz/PDF_surveillance/Lablink/

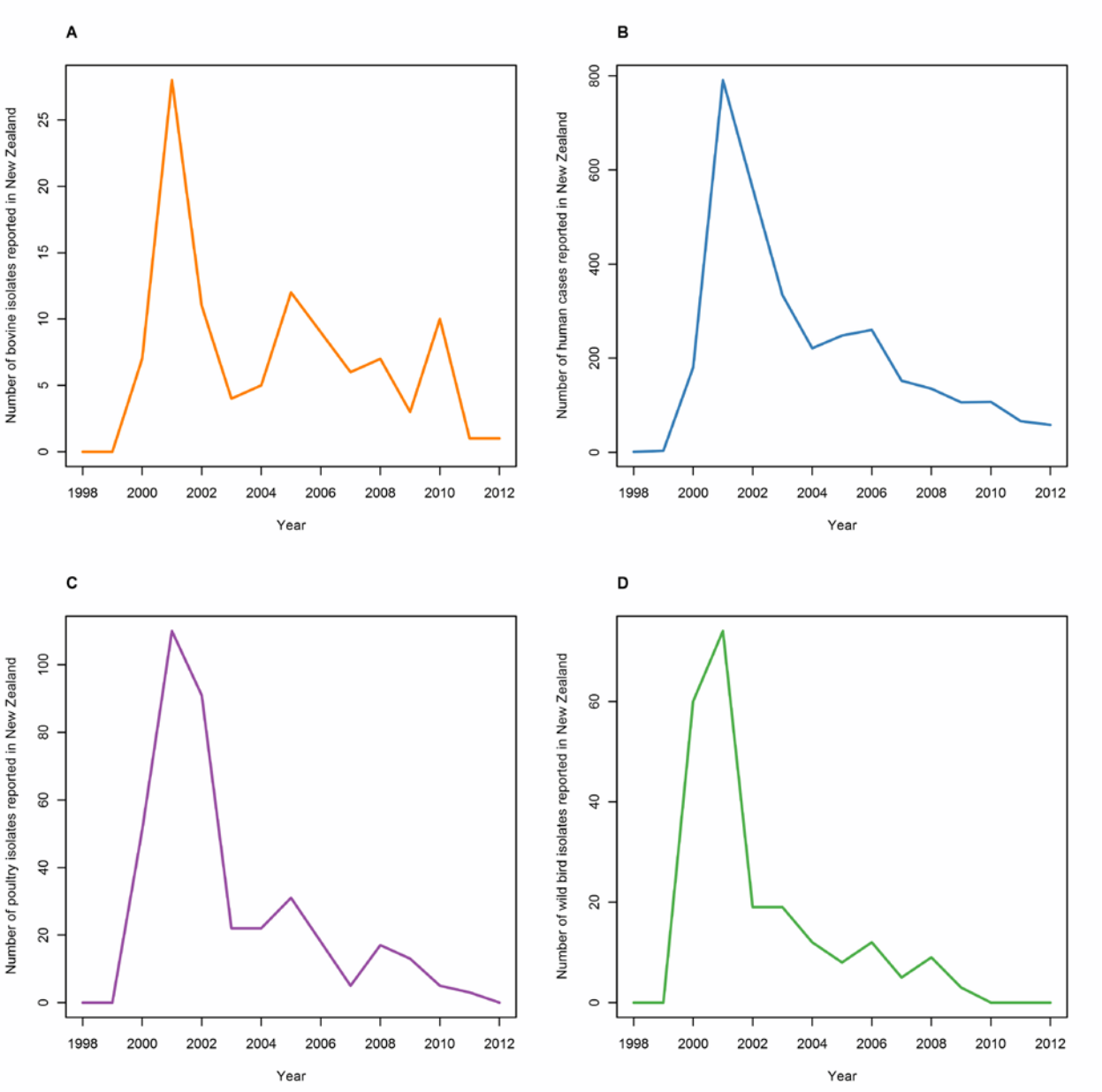
11. Institute of Environmental Science and Research Ltd (ESR). Public Health Surveillance; Information for New Zealand public health action. 2003–2012 human *Salmonella* isolates [cited 2016 Nov 25]. https://surv.esr.cri.nz/enteric_reference/human_salmonella.php

12. Institute of Environmental Science and Research Ltd (ESR). Public Health Surveillance; Information for New Zealand public health action. Non-human *Salmonella* isolates, 2003–2012 [cited 2016 Nov 25. https://surv.esr.cri.nz/enteric_reference/nonhuman_salmonella.php

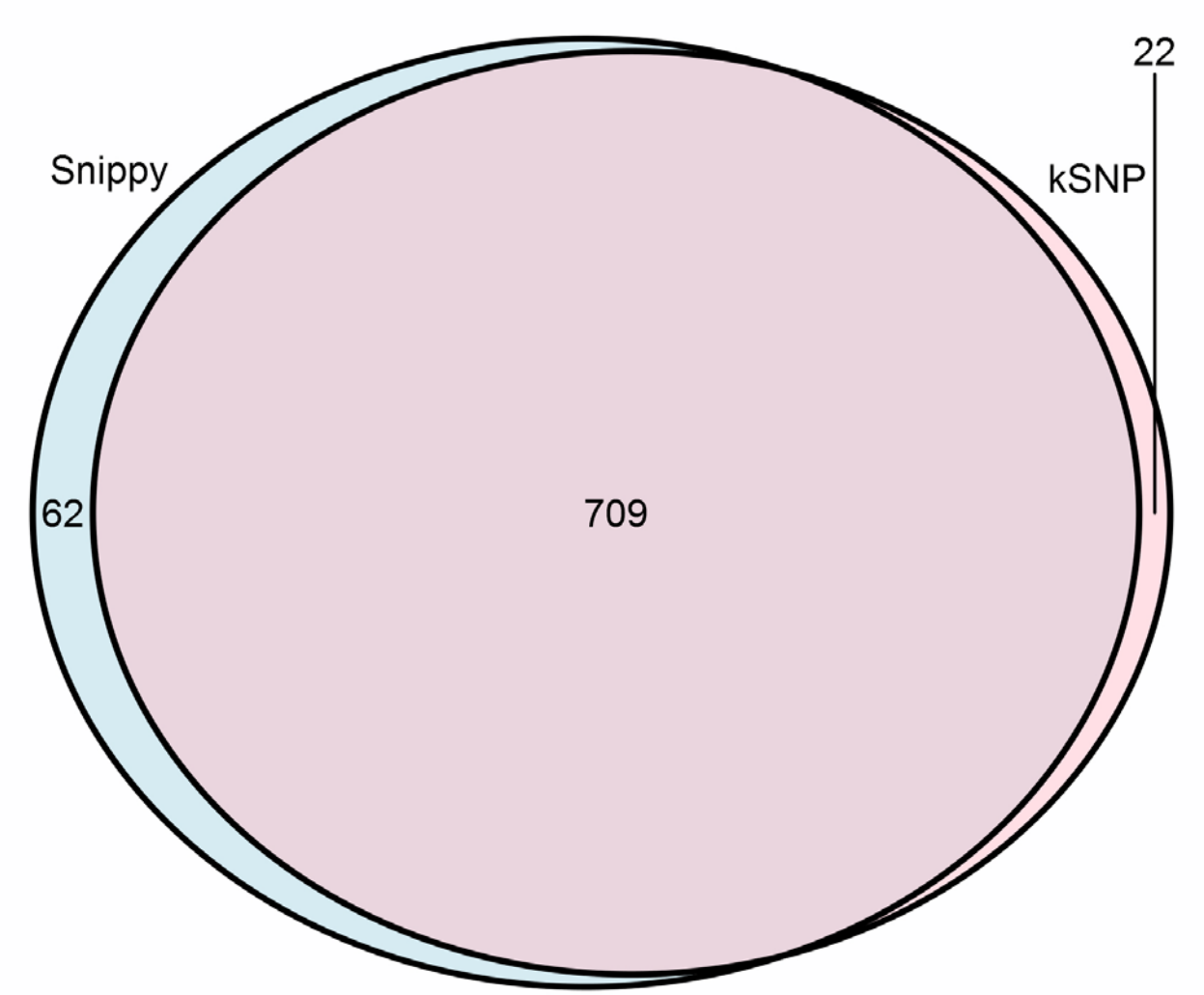
Technical Appendix Table. PERMANOVA (http://www.primer-e.com/permanova.htm) output for 107 *Salmonella enterica* serovar Typhimurium DT160 isolates, based on the presence of 684 protein differences and grouped by year of collection and source*

Coefficient	Df	SS	MSS	Pseudo-F	P(perm)	Unique perms
Year	4	42.26	10.57	1.143	0.121	998
Source	3	26.9	8.968	0.97	0.515	997
Year ×Source†	10	99.9	9.99	1.081	0.187	996
Residuals	89	822.8	9.245			
Total	106	1,002				

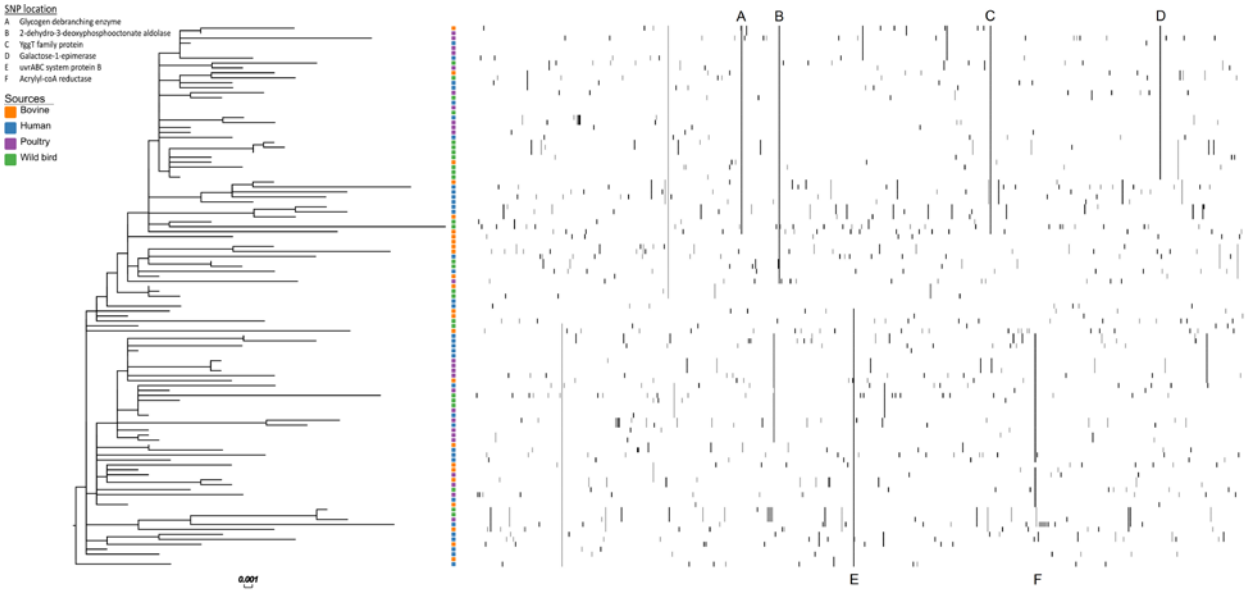
*Df, degrees of freedom; SS, sum of squares; MSS, mean sum of squares; Pseudo-F, F-value from the data; P(perm), proportion of permuted datasets whose F-value exceeds Pseudo-F; Unique perms, number of unique permutations.
†Coefficient interaction.



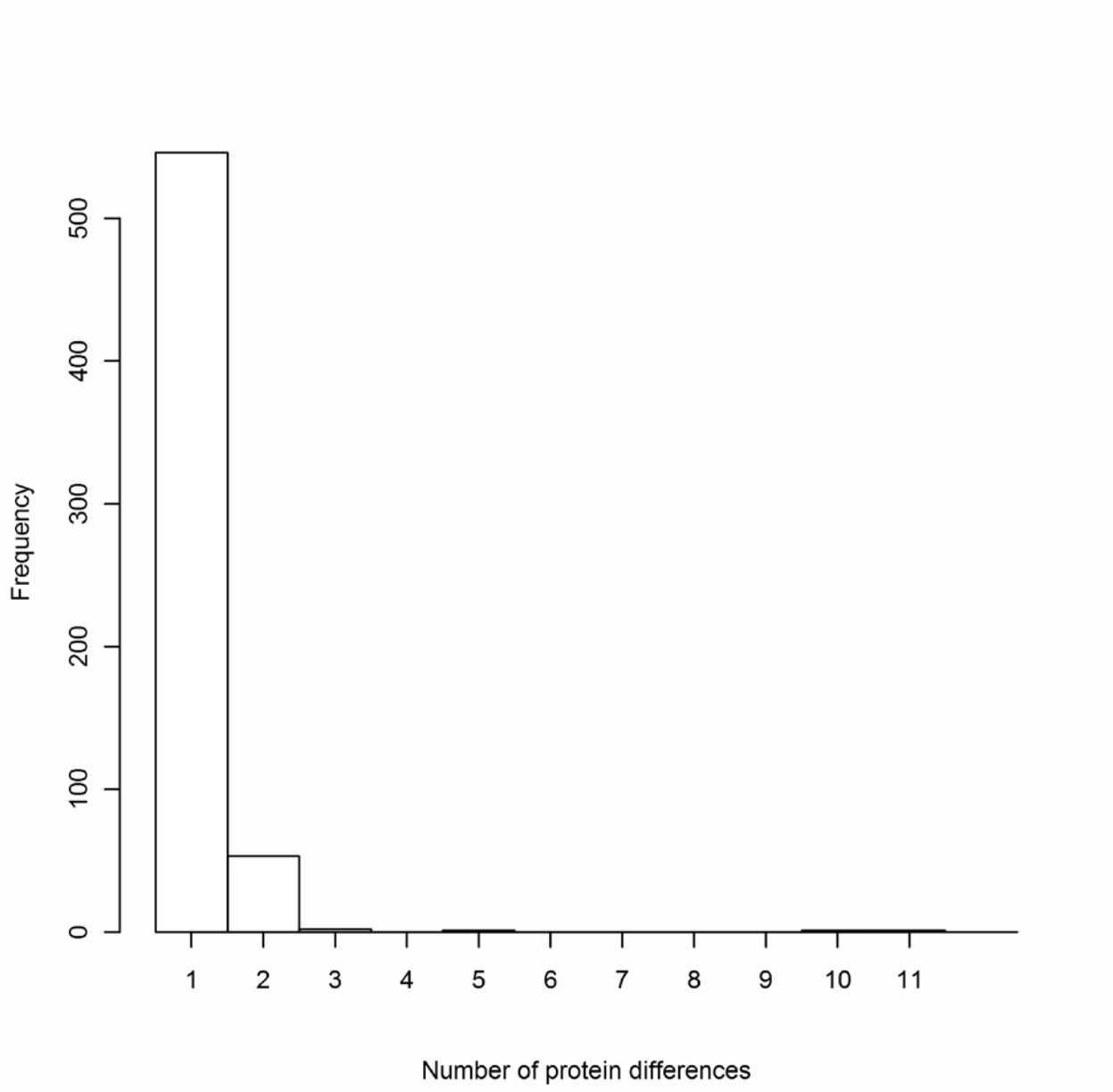
Technical Appendix Figure 1. Line graph of the number of bovine (A: orange), human (B: blue), poultry (C: purple) and wild bird (D: green) DT160 cases reported in New Zealand from 1998–2012 (*10*–*12*).



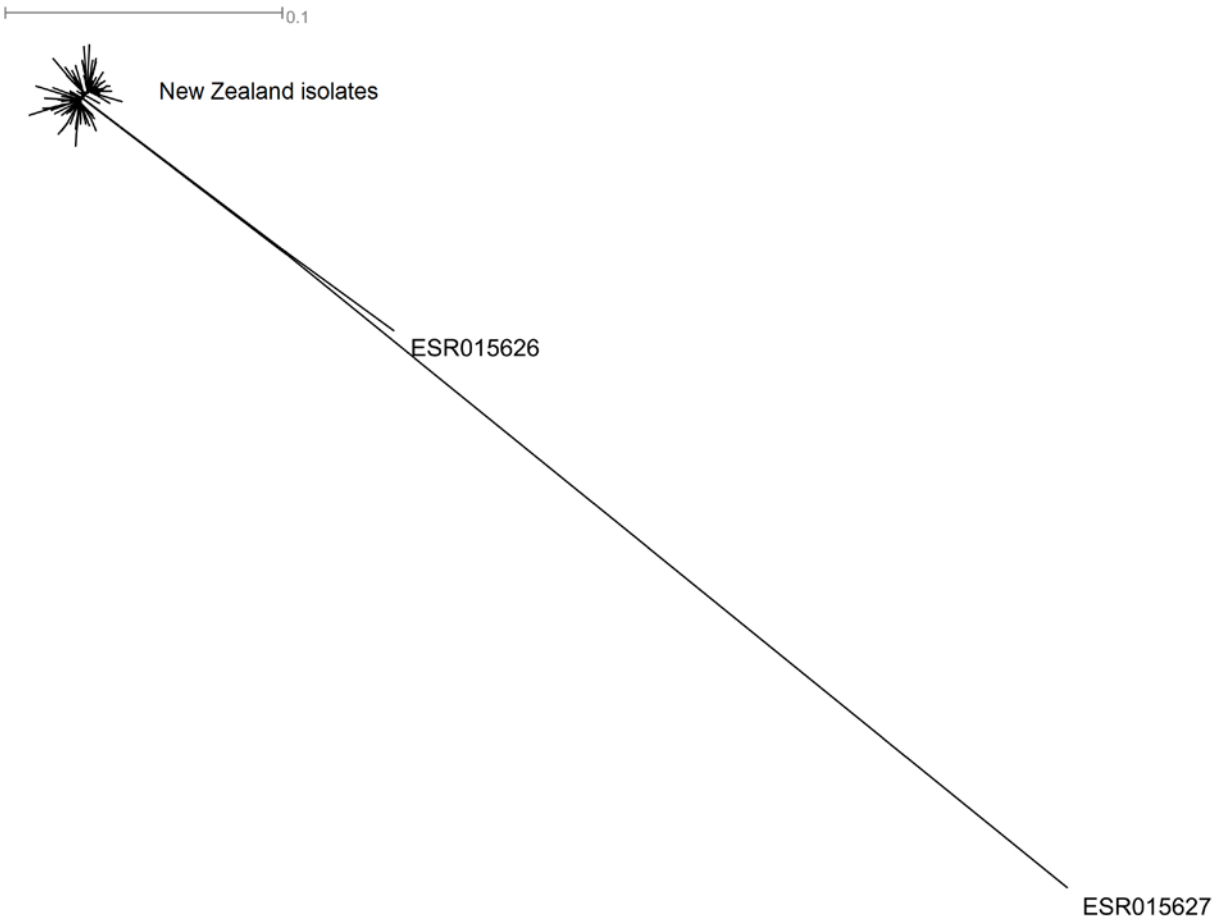
Technical Appendix Figure 2. Venn diagram of the number of unique and shared DT160 SNPs identified by Snippy and kSNP3.



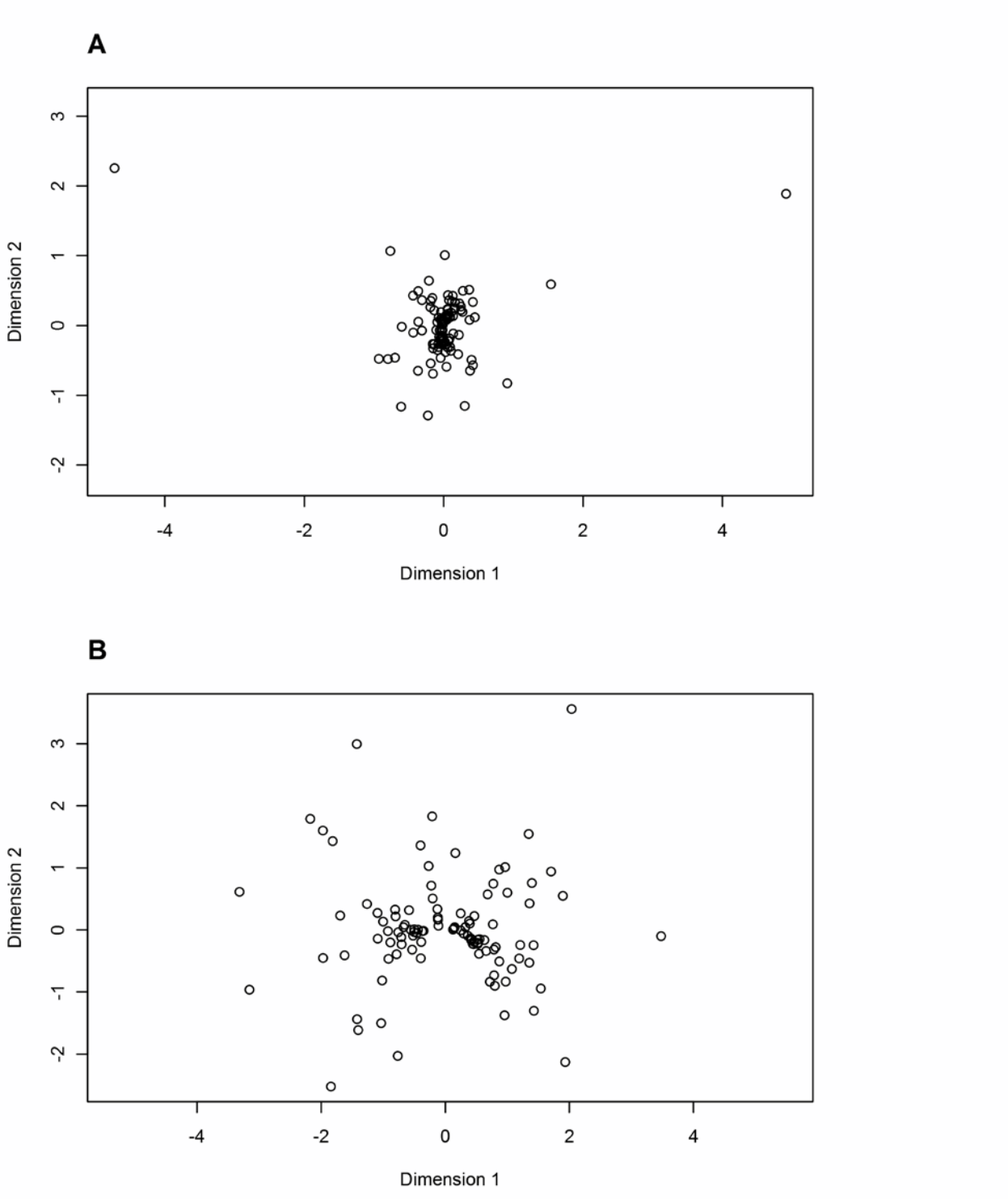
Technical Appendix Figure 3. Maximum likelihood tree of 109 DT160 isolates (based on 793 core SNPs). The scale bar represents the number of nucleotide substitutions per site. The colored squares represent the sources of the isolates. The presence-absence matrix represents the presence of the 773 core SNPs located on the reference genome, NC_016856. The SNPs were arranged in the order they appear on the reference genome. Black bars represent non-synonymous SNPs and gray bars represent synonymous SNPs. The non-synonymous SNPs responsible for the formation of the major DT160 clades were assigned a letter (A-F) and the proteins they are located within are outlined.



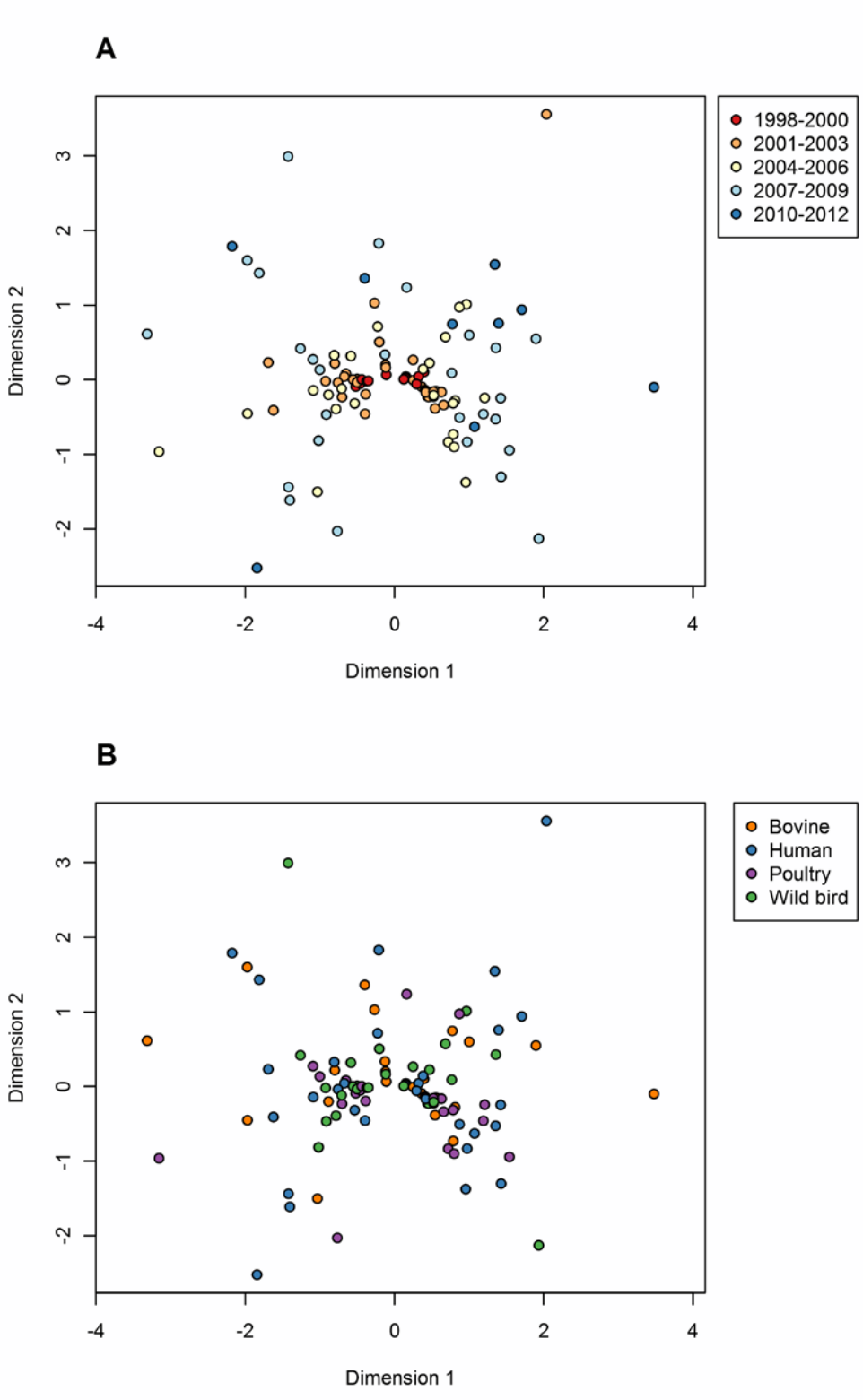
Technical Appendix Figure 4. Histogram of the number of protein differences found within the same protein sequence.



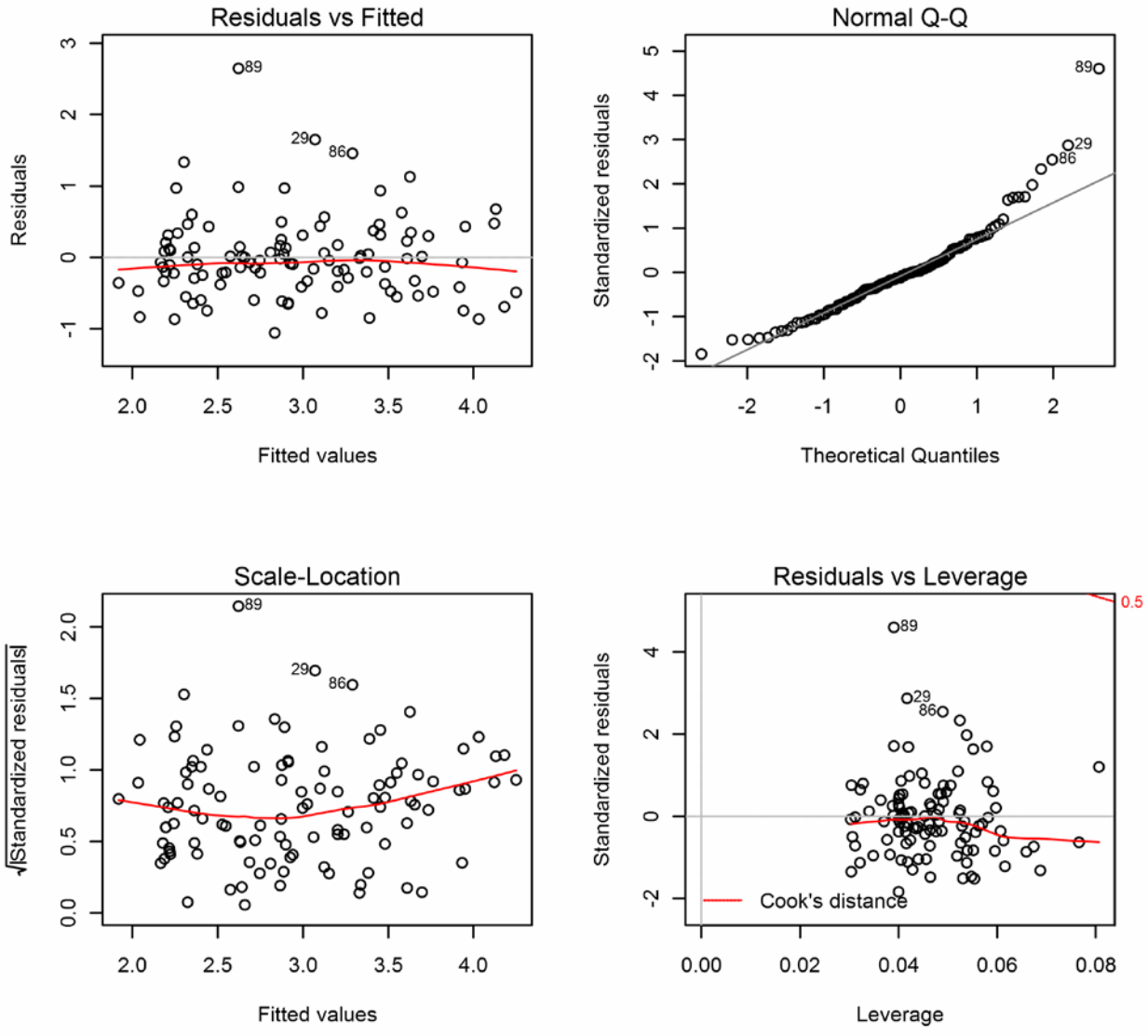
Technical Appendix Figure 5. NeighborNet tree of 111 DT160 isolates (based on 1,521 core SNPs): 109 from New Zealand and two from the United Kingdom (ERS015626 and ERS015627). The scale bar represents the number of nucleotide substitutions per site.



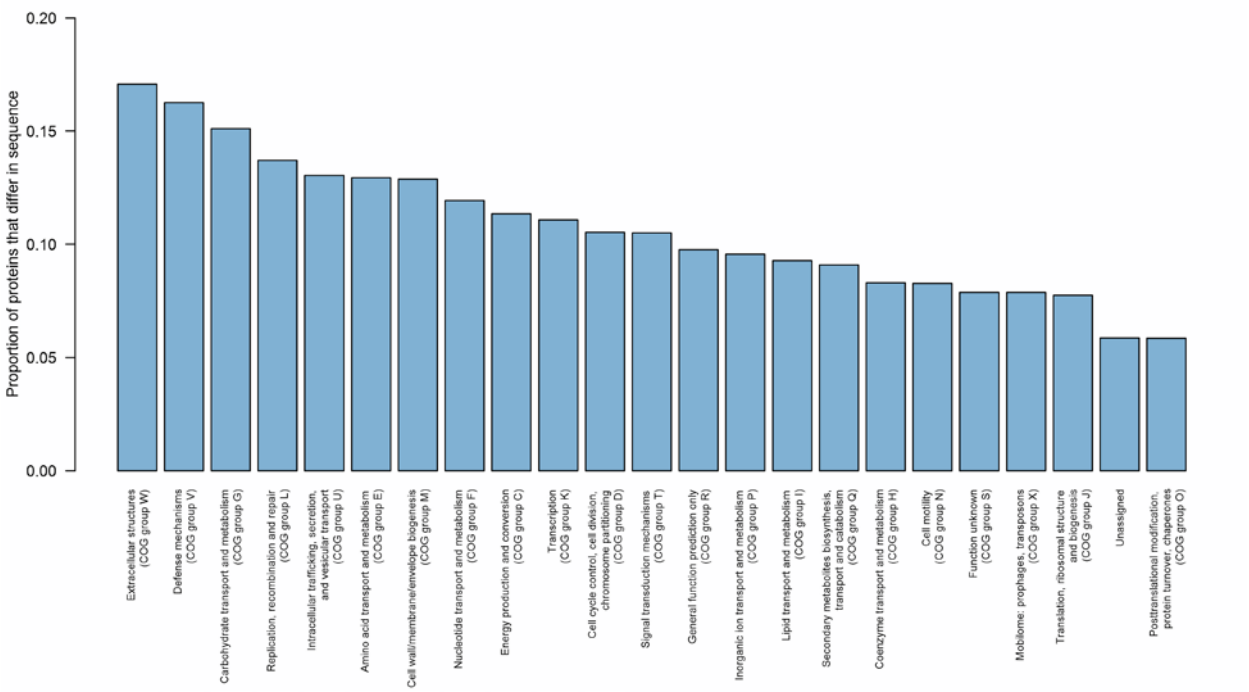
Technical Appendix Figure 6. Multi-dimensional scaling of 109 (A) and 107 (minus two outliers) (B) DT160 isolates based on the presence of 684 protein differences.



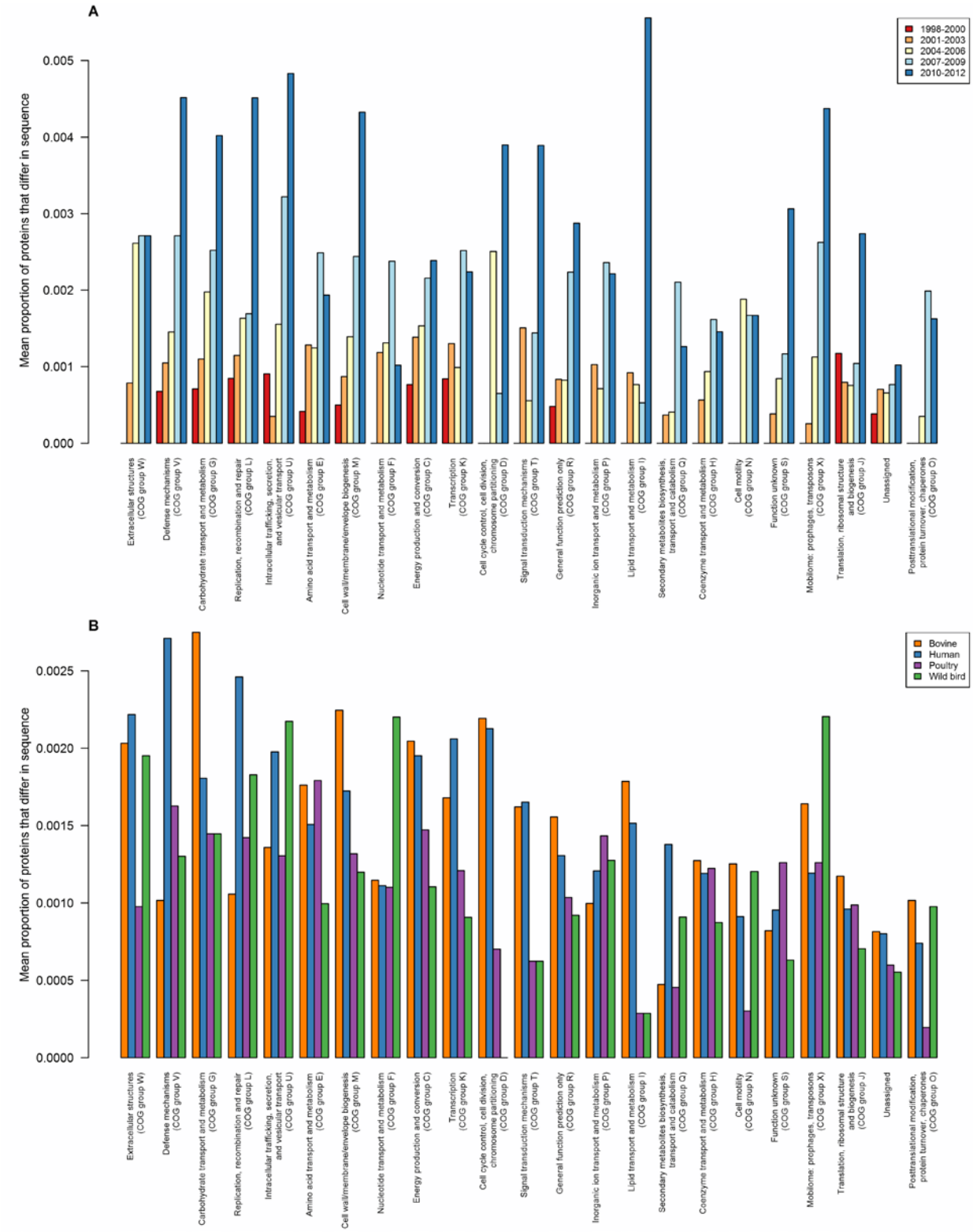
Technical Appendix Figure 7. Multi-dimensional scaling of 107 DT160 isolates, based on the presence of 684 protein differences and colored by date of collection (A) and source (B).



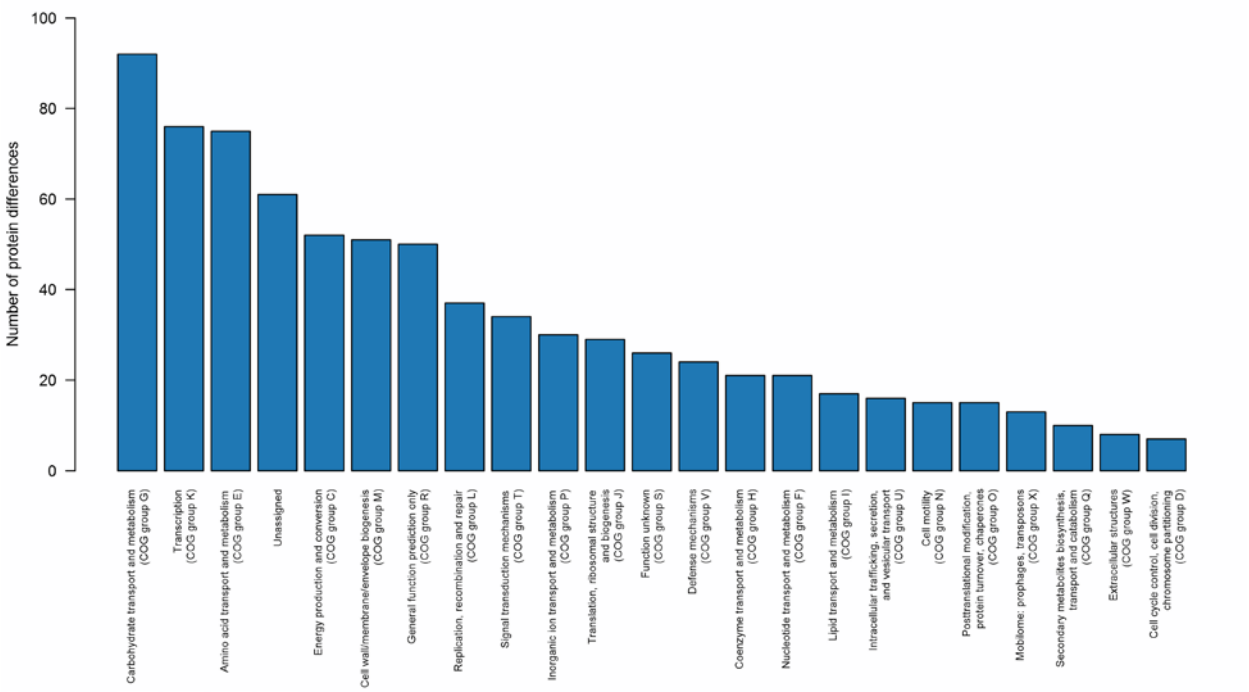
Technical Appendix Figure 8. Diagnostic plots of the regression model fitted to the z-values for 107 DT160 isolates.



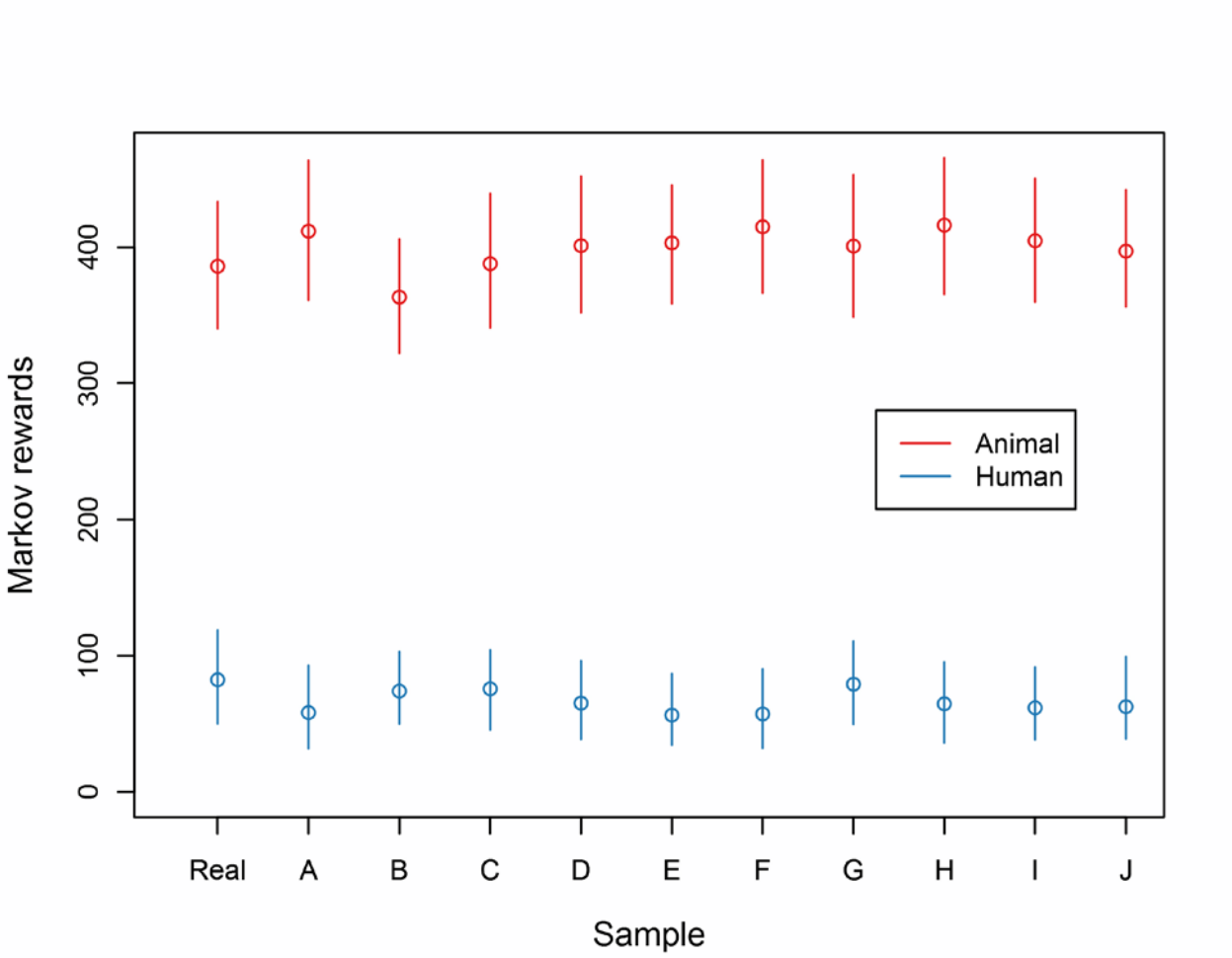
Technical Appendix Figure 9. Bar graph of the proportion of proteins that differ in sequence for each COG functional group.



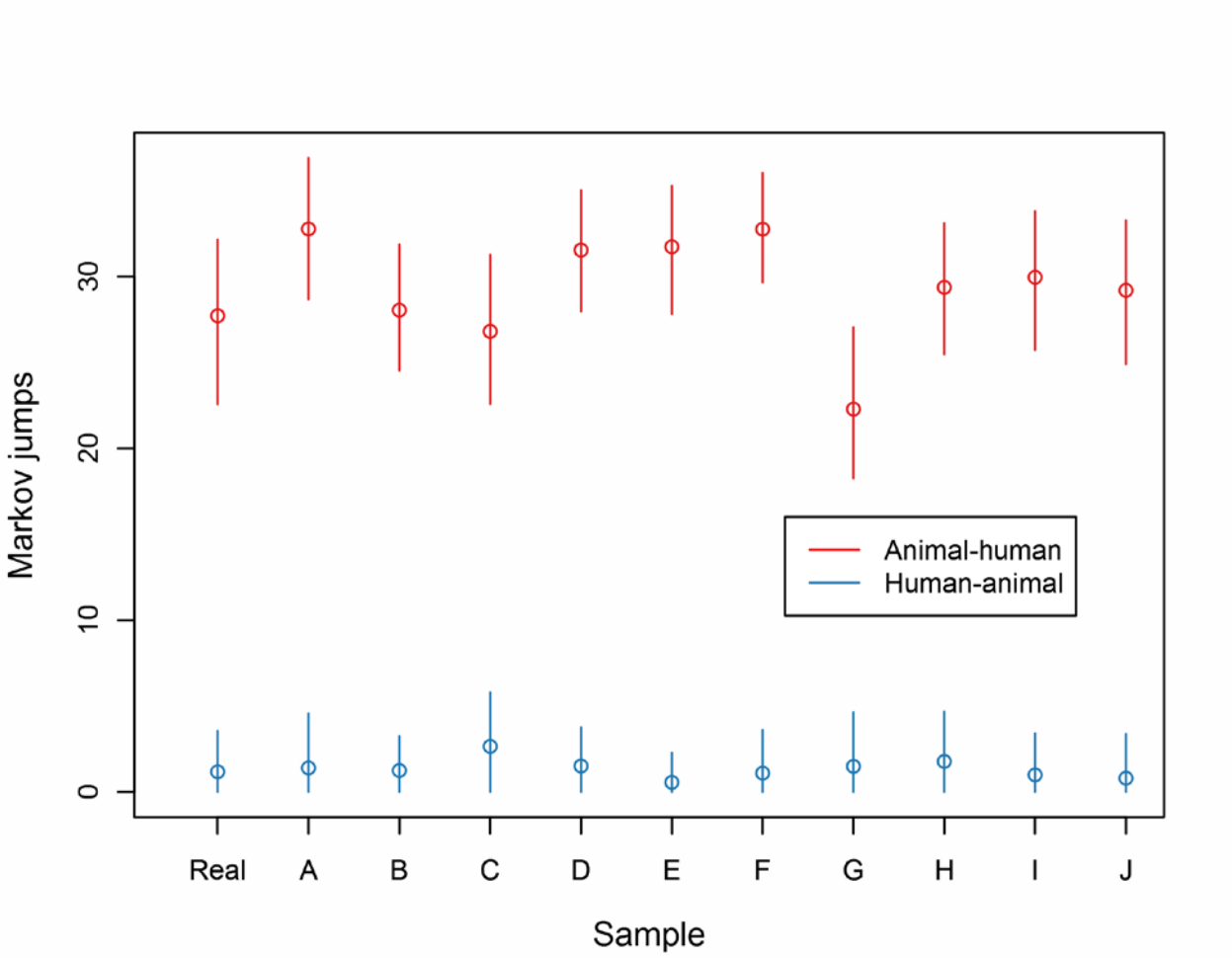
Technical Appendix Figure 10. Bar graph of the mean proportion of proteins that differ in sequence for each COG functional group within each time period (A) and source (B).



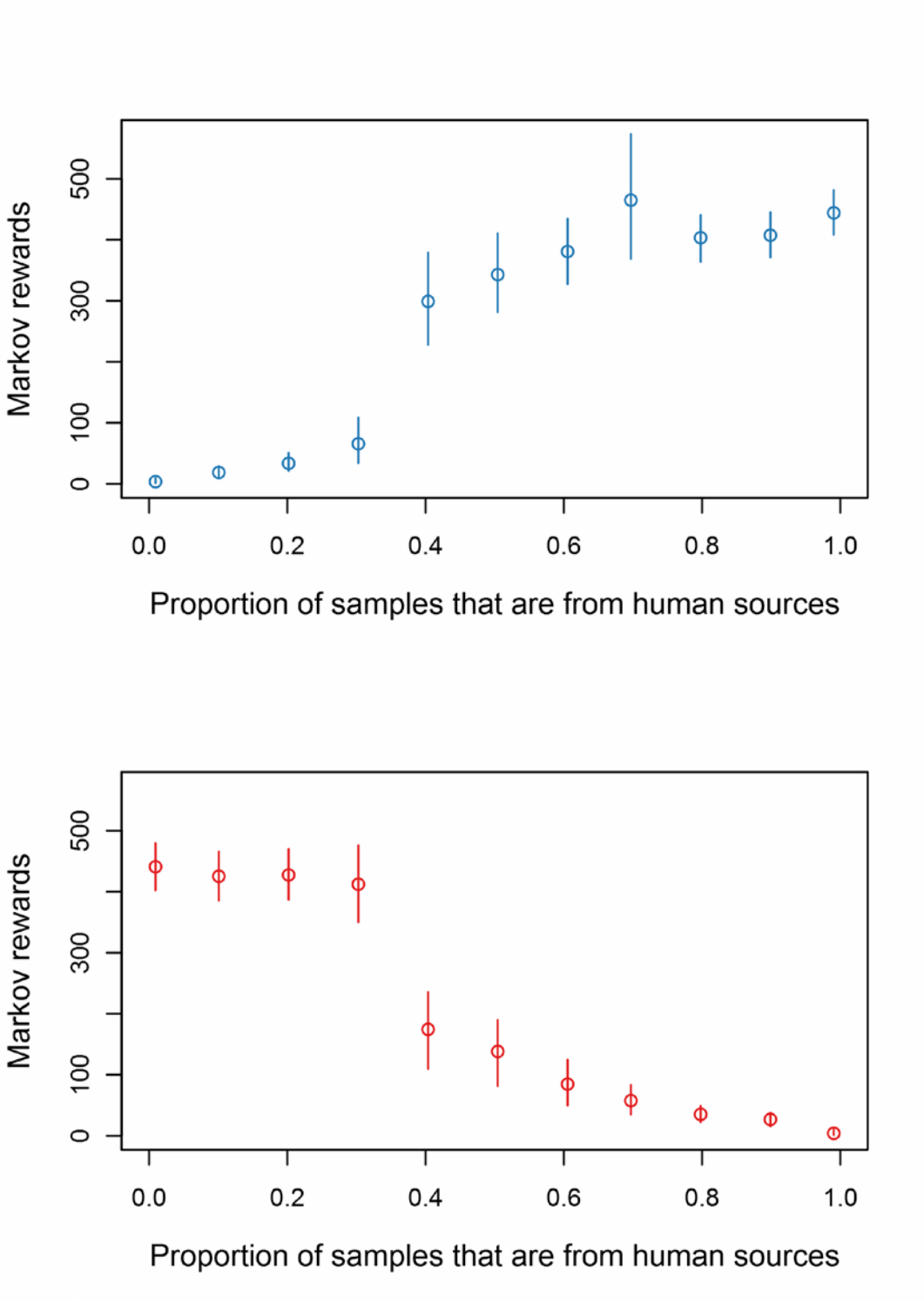
Technical Appendix Figure 11. Bar graph of the number of protein difference for each functional group shared by 107 DT160 isolates.



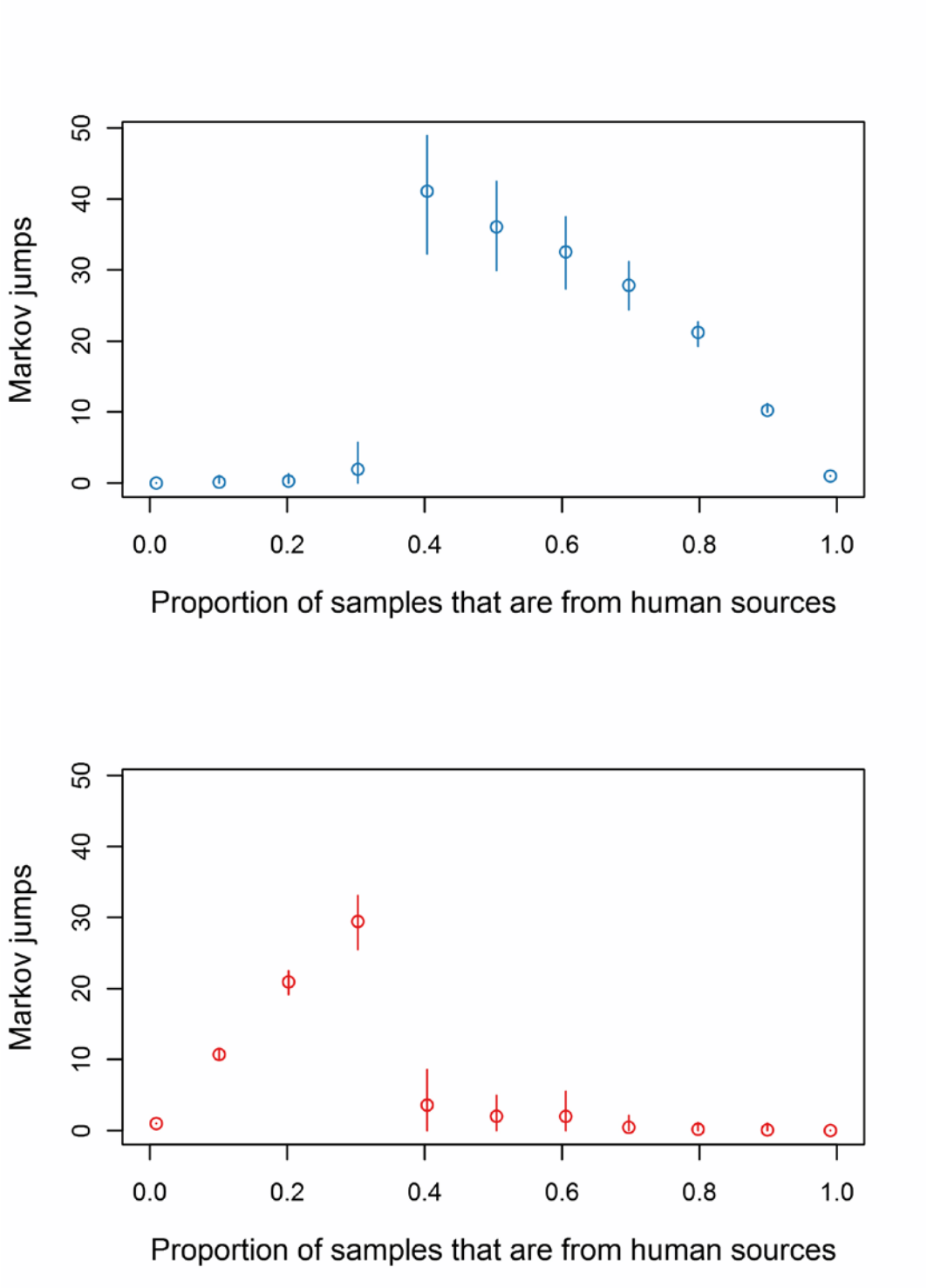
Technical Appendix Figure 12. Scatter plot of the number of animal (red) and human (blue) Markov rewards estimated for the real and ten randomly assigned (A-J) datasets. The circles represent the mean Markov reward value and the error bars represent the 95% HPD interval.



Technical Appendix Figure 13. Scatter plot of the number of animal-to-human (red) and human-to-animal (blue) Markov jumps estimated for the real and ten randomly assigned (A-J) datasets. The circles represent the mean Markov reward value and the error bars represent the 95% HPD interval.



Technical Appendix Figure 14. Scatter plot of the number of animal (blue) and human (red) Markov rewards estimated versus the proportion of samples assigned as human. The circles represent the mean Markov reward value and the error bars represent the 95% HPD interval.



Technical Appendix Figure 15. Scatter plot of the number of animal-to-human (blue) and human-to-animal (red) Markov jumps versus the proportion of samples assigned as human. The circles represent the mean Markov jump value and the error bars represent the 95% HPD interval.

Technical Appendix Figure 16. Maximum clade credibility trees of 109 DT160 isolates placed through the discrete phylogeographic model, with different proportions of isolates assigned as human (blue) and animal (red).