Support Vector Machine applied to predict the zoonotic potential of E. coli O157 cattle isolates

Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates

Nadejda Lupolova[a], Timothy J. Dallman[b], Louise Matthews[c], James L. Bono[d], and David L. Gally[a,1]

[a]Division of Immunity and Infection, The Roslin Institute and The Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, United Kingdom; [b]Public Health England, National Infection Service, London NW9 5EQ, United Kingdom; [c]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom; and [d]US Meat Animal Research Center, Agricultural Research Service, United States Department of Agriculture, Clay Center, NE 68933

Edited by Roy Curtiss III, University of Florida, Gainesville, FL, and approved August 2, 2016 (received for review May 2, 2016)

Sequence analyses of pathogen genomes facilitate the tracking of disease outbreaks and allow relationships between strains to be reconstructed and virulence factors to be identified. However, these methods are generally used after an outbreak has happened. Here, we show that support vector machine analysis of bovine *E. coli* O157 isolate sequences can be applied to predict their zoonotic potential, identifying cattle strains more likely to be a serious threat to human health. Notably, only a minor subset (less than 10%) of bovine *E. coli* O157 isolates analyzed in our datasets were predicted to have the potential to cause human disease; this is despite the fact that the majority are within previously defined pathogenic lineages I or I/II and encode key virulence factors. The predictive capacity was retained when tested across datasets. The major differences between human and bovine *E. coli* O157 isolates were due to the relative abundances of hundreds of predicted prophage proteins. This finding has profound implications for public health management of disease because interventions in cattle, such a vaccination, can be targeted at herds carrying strains of high zoonotic potential. Machine-learning approaches should be applied broadly to further our understanding of pathogen biology.

machine learning | zoonosis | Shiga toxin | *E. coli* | cattle

For important global bacterial zoonoses such as *Salmonella*, enterohemorrhagic *Escherichia coli* (EHEC), and *Campylobacter*, tracking of disease outbreaks and identification of infection source are critical to limiting further disease. Whole-genome sequencing (WGS) has provided a revolution in our capacity to identify and trace outbreaks that would have been virtually impossible with more traditional techniques such as phage typing and pulsed-field gel electrophoresis (1, 2). Currently, most analyses rely on extraction of a core "shared" genome and isolate relationships are deduced based on SNPs in this core; conversely, accessory genome information is largely ignored due to its variability, although a number of approaches have recently been applied to interrogate pan-genome data (3).

EHEC infections, in particular by serogroups O157 and O26 (4), have emerged as a serious threat to human health in the last 30 y, driven by the integration of bacteriophages encoding Shiga toxin (Stx) into the genomes of specific *E. coli* strain backgrounds. Strains encoding Stx subtype 2a and a type 3 secretion system are often associated with the most severe human infections, which can lead to bloody diarrhea (hemorrhagic colitis) and kidney damage. Stx kills capillary endothelial cells and the host's attempt to repair this damage can result in red blood cell hemolysis in capillaries known as hemolytic uremic syndrome, which can be fatal (5–7). There has been extensive work to determine which strains in ruminants, in particular cattle, represent the most serious threat to human health (6, 8, 9). This led to the definition of lineages and clades for which lineage I or lineage I/II are more likely to be associated with human disease, whereas lineage II strains are more restricted to cattle (10, 11). Within these lineages certain clades predominate, so clade 8 within lineage I/II has been associated in the United States with more serious disease in humans (12). In the United Kingdom, a recent WGS analysis of over 1,000 EHEC O157 human and cattle isolates was used to determine their phylogeny based on core genome SNP analysis (13). The most serious disease in the United Kingdom is associated with lineage I strains and a specific phage type (PT) designated PT21/28; phage typing of UK strains is based on susceptibility testing with a collection of diagnostic bacteriophages (14). The United Kingdom has a high incidence of serious EHEC O157 infections, and the emergence of these infections in the 1990s coincided with the acquisition of the Stx 2a subtype into UK cattle strains already encoding a Stx2c subtype (13, 15).

Current core genome analysis of EHEC strains indicates complete mixing of human and bovine EHEC O157 isolates (Fig. 1*A* and Fig. S1). This fits with the concept that the majority of cattle strains within particular lineages and encoding Stx 2a are a serious threat to human health. In the present study, we aimed to determine whether a pan-genome analysis of EHEC O157 strains could distinguish between human isolates and isolates from cattle. In particular, we wanted to test whether machine-learning approaches such as support vector machine (SVM) (16) could be used to discriminate a subset of bovine strains that might represent a threat to human health and would allow more targeted interventions in cattle. SVM has been applied in many areas of bioinformatics, including prediction of protein function, prediction of transcription initiation site, and classification of gene expression data as well as cancer prediction and prognosis (17, 18).

Results and Discussion

UK Dataset. We initially analyzed an extensive UK dataset that consisted of WGS for 185 *E. coli* O157 strains isolated from

Significance

Zoonotic infections with enterohemorrhagic *Escherchia coli* O157 have emerged as a serious threat to human health. Conventional sequence-based analyses indicate that most human infections originate from particular pathogenic lineages. In this study, we apply a machine-learning approach to complex pangenome information and predict the human infection potential of cattle *E. coli* O157 isolates. We demonstrate that only a small subset of bovine strains is likely to cause human disease, even within previously defined pathogenic lineages. The approach was tested across isolates from the United Kingdom and United States and verified with food and cattle isolates from outbreak investigations. This finding has important implications for targeting of control strategies in herds.

Author contributions: N.L., T.J.D., L.M., and D.L.G. designed research; N.L. performed research; N.L., T.J.D., and J.L.B. contributed new reagents/analytic tools; N.L. and D.L.G. analyzed data; and N.L., T.J.D., J.L.B., and D.L.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: dgally@ed.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606567113/-/DCSupplemental.

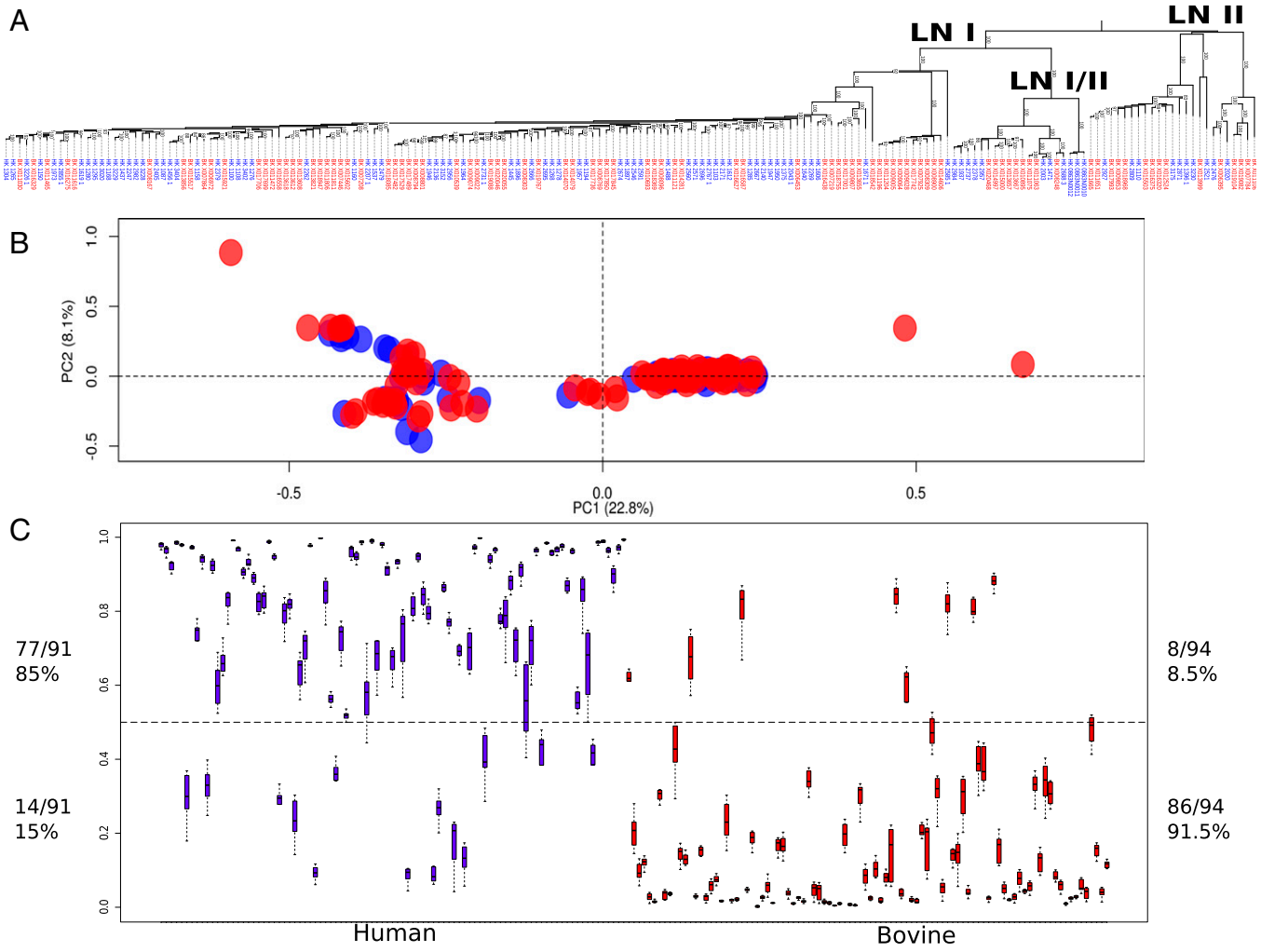

Fig. 1. UK Shiga toxin-producing *E. coli* (STEC) O157 dataset analysis. (*A*) Core SNP phylogenetic tree. The three main lineages (LN) are shown. The majority of the UK isolates are in lineage I (LN I) with bovine (red) and human (blue) isolates interspersed across the tree. Bootstrap values shown on branches. (*B*) MMDS plot with each isolate represented by a circle. The denser cluster on the right-hand side is composed primarily of LN I isolates with equivalent numbers of human isolates overlaid by bovine isolates. (*C*) SVM probability plot based on repeated testing of isolates in the different subsets. The probability of each isolate belonging to the human or bovine group was calculated over random repeated samples; median values are shown with interquartile ranges. The predicted "host" of the isolate is based on whether the mean probability is below 0.5 (bovine) or above (human). The percentages of isolates assigned to each host by the model are shown at the sides of the graph.

human patients in the United Kingdom ($n = 91$) and cattle ($n = 94$). The 185 strains share 4,737,622 core positions, which is equivalent to 85% of the reference *E. coli* Sakai strain genome (19). A maximum likelihood phylogenetic tree based on these positions clearly splits into distinctive branches, even within this relatively clonal serotype (Fig. 1*A*). The pattern for the UK O157 phylogenetic tree is consistent with previous studies (11, 13, 20) and represents a typical split for UK strains based on lineages: lineage I ($n = 140$, 70 bovine), II ($n = 25$, 15 bovine), and I/II ($n = 17$, 9 bovine). The average number of SNPs within two sequences of the same lineage was 1,859, 379, and 2,190 for lineages I, I/II, and II, respectively. The vast majority of the lineage I sequences were PT21/28 (101 out of 140) and the second most prevalent was PT32 (24 out of 140). The dominant PT in lineage II was PT8 (15 out of 25) and in lineage I/II was PT2 (14 out of 17). Based on phylogenetic analysis of core SNPs, it was not possible to detect any evidence of clustering by human or bovine host (Fig. 1*A*).

Determination of the accessory genome using the Roary pan-genome pipeline indicated that among 185 UK isolates there were 14,636 protein clusters assigned, based on 95% amino acid sequence similarity. Core proteins present in more than 95% of the sequences generated 4,369 clusters; 979 clusters originated from proteins predicted in 15–95% of sequences, leaving a high number of rare clusters (9,288) that were present in less than 15% of isolates. The majority of all protein clusters (10,653) were annotated (i.e., were similar to already-annotated proteins from a public database) and 3,983 were hypothetical. There were only 5,485 unique protein names across all of the genomes, and 3,807 of these produced single copy clusters. Due to these rules of cluster assignments, many homologous proteins generated multiple clusters. We have termed these protein variant (PV) clusters. An exceptionally high number of PVs were produced by phage-related proteins, confirming that phage sequences are highly variable (21).

An accepted way to analyze complex pan-genome data is to apply metric multidimensional scaling (MMDS) with different methods of matrix distance calculations. In the present study, methods of distance calculation had little effect on the final MMDS plots, and thus all MMDS plots presented in this paper are based on simple dissimilarity calculations (Fig. 1*B*). Dense clusters on the MMDS plot were highly correlated to the lineages shown on the phylogenetic tree. Thus, further clustering of UK isolates by *k*-means resulted in two clusters: one with all isolates having 100% support and originating from lineage I (128 isolates

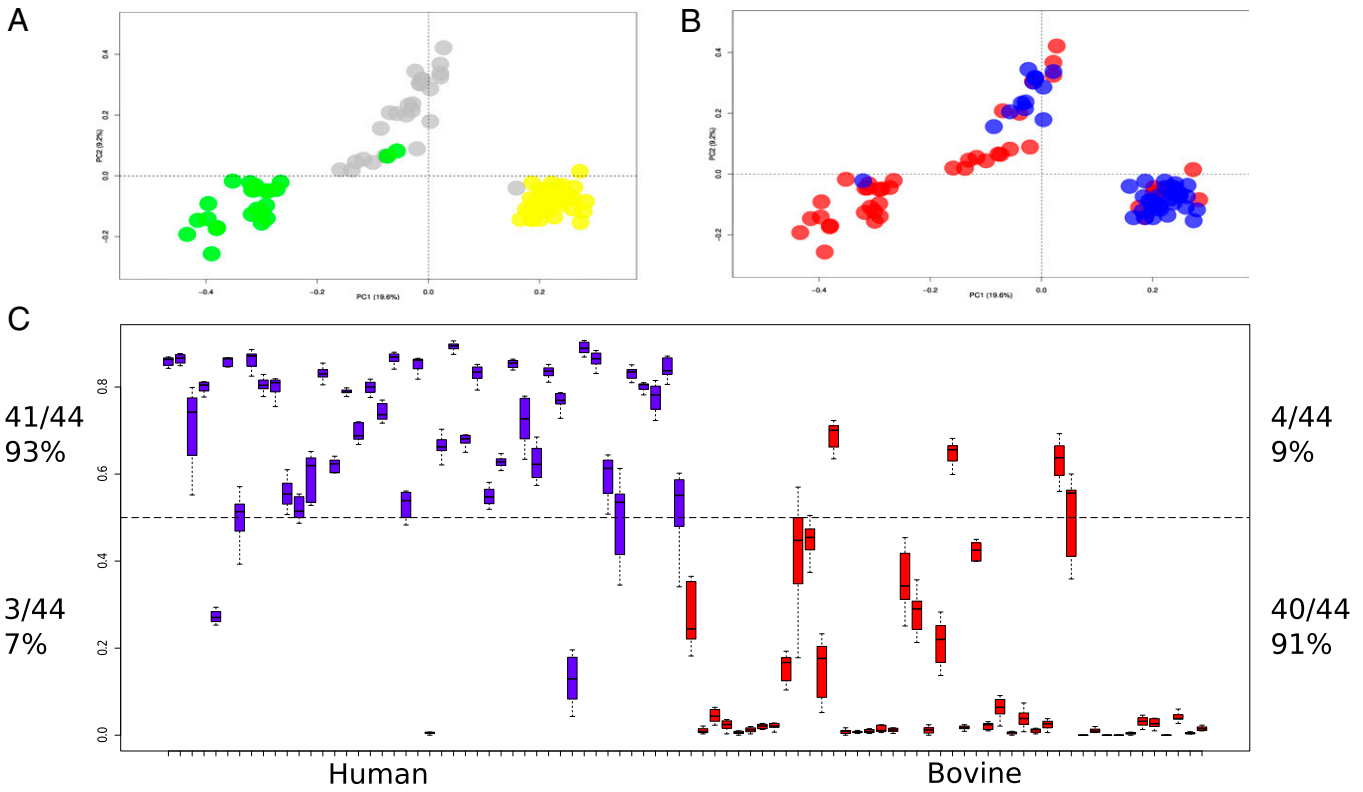

Fig. 2. US STEC O157 dataset analysis. MMDS analysis of US pan-genome dataset with each isolate represented by a circle. (*A*) MMDS clustering with isolates colored by lineage: lineage I in yellow, lineage I/II in gray, and lineage II in green. (*B*) MMDS clustering with isolates colored by host: red, bovine isolates and blue, human isolates. (*C*) SVM probability plot based on repeated testing of isolates in the different subsets. The probability of each isolate belonging to the human or bovine group was calculated over random repeated samples with median values and interquartile ranges shown. Red shading for bovine isolates; blue shading for human isolates. The predicted "host" of the isolate is based on whether the mean probability is below 0.5 (bovine) or above (human). The percentages of isolates assigned to each host by the model are shown at the sides of the graph.

out of all 140 lineage I isolates) and the second with isolates primarily from lineages II and I/II (support higher than 90%) and with only 11 isolates from lineage I (support between 85–90%). All isolates in the second cluster that belonged to lineage I were PT32 (Fig. S2). Overall, MMDS methods provided results similar to the phylogenetic analysis, namely, separation of lineage I and little capacity to distinguish between human and bovine isolates.

Machine learning methods have been routinely applied to investigate complex data in several areas of science, although, until now, it has not been used to analyze bacterial genomic data to predict phenotype from the genotype. Therefore we built an SVM classifier trained on *E. coli* isolates with known isolation host (human or cattle) and tested whether the classifier could predict the likely host origin (human/bovine) of isolates from their PV profile. To choose the features for the model, the proportions of each PV present in each host group were calculated separately. There were a total of 10,878 clusters with a different proportion of PVs between the two hosts (Table S1). To reduce the number of features introduced into our model, while preserving accuracy, we used only PVs with a subtractive difference between the two hosts of >10 (n = 638) and have defined the discriminatory PVs at >20 (n = 82) in Tables S2 and S3. The probability of each isolate being assigned to the human or bovine group was then calculated by random repeated sampling and the resulting probabilities plotted in Fig. 1*C*. Overall, using a probability of 0.5 as the separation value, 85% of human and 91% of bovine isolates were assigned in accordance with the host from which they were isolated, and the majority with high probabilities. This shows that it is possible to differentiate these isolates based on the isolation host, indicating that host-specific information for *E. coli* O157 can be derived from the sequence data alone. Because ruminants, in particular cattle, are a primary reservoir for EHEC O157 strains, there was an a priori assumption that it may not be possible to assign isolates to the two host groups because the majority of human isolates are likely to originate from cattle. However, this was not the case, and it is an important observation that a minor subset of isolates originating from cattle were classified into the human group (Fig. 1*C*). These same bovine isolates were persistently called as human, meaning that the model does find features in these isolates that make them more similar to those from the human population than from cattle. This finding indicates that not all bovine isolates have the same zoonotic potential; in fact, the majority of bovine *E. coli* O157 isolates were not predicted to be associated with human disease.

The majority of either bovine or human isolates did not change their assignment probabilities with multiple subtesting (the majority close to 0 or 1) and strains called distinct from their isolation host were called so consistently. Midrange isolates had more variable assignment probabilities (Fig. 1*C*) and this may indicate genomes with both human- and bovine-specific features. In addition, the bovine isolates called as "human" and the isolates called in the reverse direction cannot be explained by available metadata including lineage and PT; for example, the bovine isolates represent a mixed group of PTs: PT21/28 (n = 4) and one of each PT 31, 32, 33, and 49. Six of these isolates possessed *stx2a*/*2c*, one 2a, and one was negative for *stx*. We note that MMDS analysis of this differential PV subset did not separate strains by isolation host with clustering still tied to lineages and SNP core phylogeny (Fig. S3).

SVM models can be analyzed for accuracy and prediction capacity (Fig. S4), with accuracy calculations based on the level of "incorrect" assignments. However, there is an expectation that

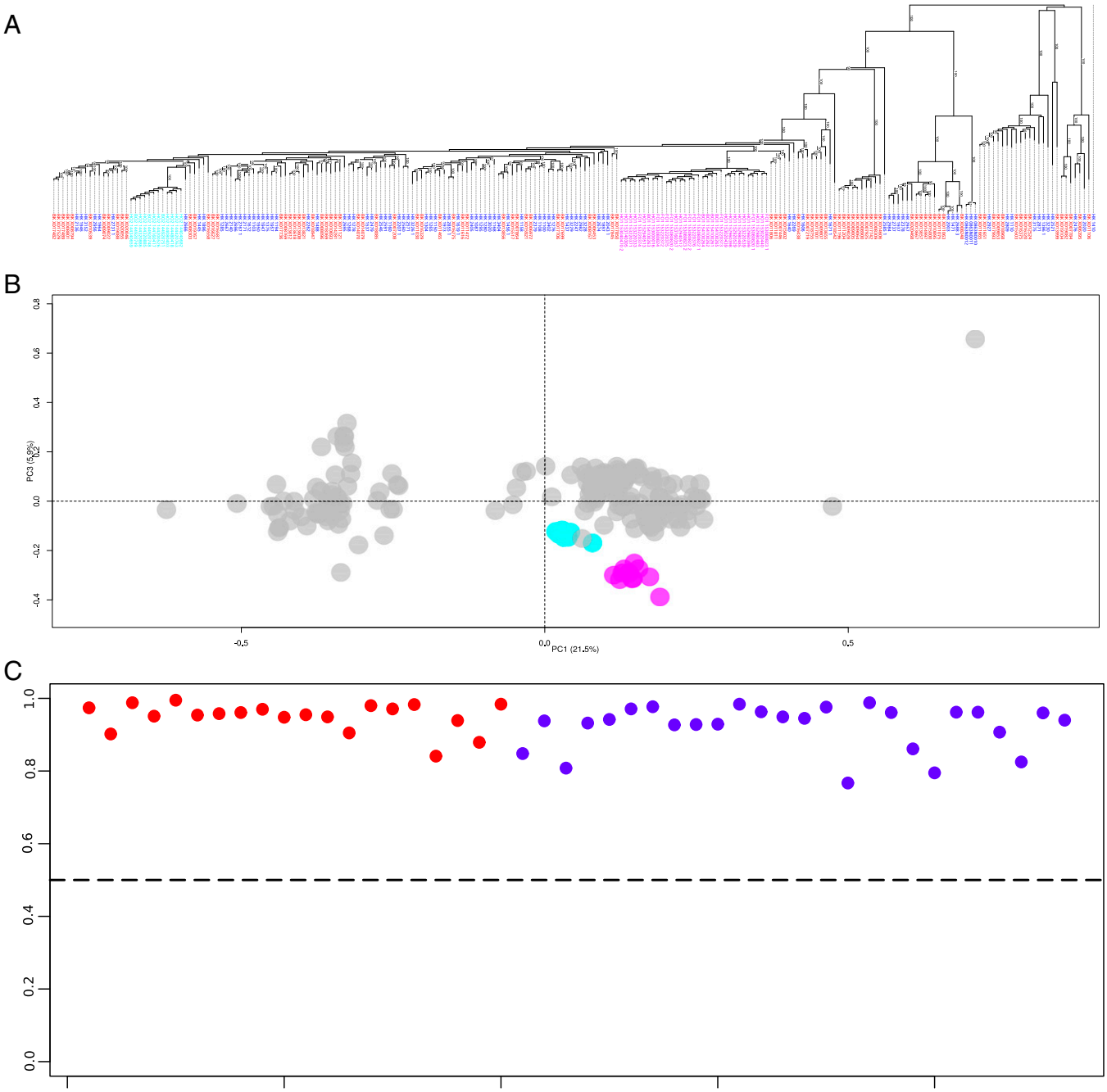

Fig. 3. Analysis of STEC O157 outbreak isolates. (*A*) Core SNP phylogenetic analysis of bovine UK (red) and human UK (blue) isolates and the two EHEC O157 outbreaks (magenta and cyan) showing that isolates from both outbreaks fall within lineage I and cluster tightly. (*B*) MMDS analysis of the two outbreaks (magenta and cyan) relative to the UK isolate subset (gray). The outbreak isolates form distinctive clusters although associated with lineage I. (*C*) SVM probability plot of each outbreak isolate without repeated sampling. Isolates from cattle, milk, or hamburger meat from both outbreaks are in red, and isolates from human hosts from both outbreaks are colored in blue. All isolates from both outbreaks (milk, hamburger, cattle, and human) were predicted to be "human."

our two host groups are not mutually exclusive, in other words that some isolates can colonize both hosts and therefore will contribute to model "inaccuracy." A logical extension of this point is that if the model were 100% accurate, then there would be no strain cross-over between the groups, indicating complete host adaptation or a very rare subset of cattle isolates with zoonotic potential. Therefore, accuracy estimations can reflect the underlying biology of the isolates and should be considered minimum estimates.

An important potential concern for data analysis by SVM is overfitting, for which the model is not using biologically relevant information to separate the groups. There are a number of ways to test for this; the most rigorous is to train the model on one dataset and then test it on a completely separate dataset. We apply this model successfully in the next section using isolate sequences from the United States. In addition, for our UK dataset we also tested whether we could train the model on two randomly labeled sets (containing both human and bovine isolates) and determined whether strains from these random groups could then be correctly assigned back to these groups. This was carried out in two ways. The first involved subsampling from our groups (random or bovine/human) with differential PVs (>10)

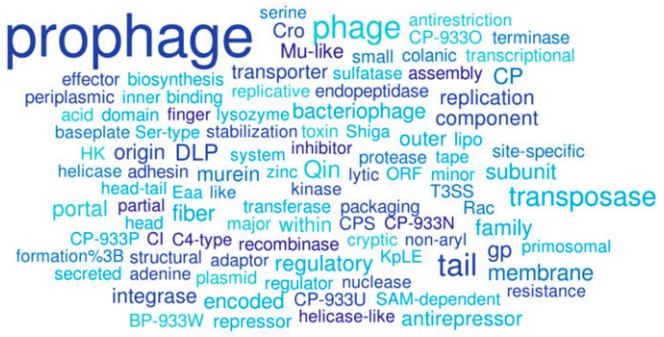

Fig. 4. Word cloud depiction of annotated PVs that were in higher proportions of human strains compared with cattle strains based on analysis of the UK dataset. Many of the PVs from human and cattle isolates are of bacteriophage origin.

determined for these subsamples. Forty isolates distinct to the training sets but within the assigned groups were then tested. This subsampling, PV determination, and testing was repeated 20 times. As expected, the isolates from the random groups had a normal distribution of probabilities reflecting their random assignment (Fig. S5*C*); in contrast, the bovine/human isolates had a different distribution with the majority having high probabilities of host assignment (i.e., 1 and 0) (Fig. S5*A*). Moreover, the repeated subsampling of the bovine/human groups yielded a significantly higher mean number of PVs (637.1, SD = 63, SE = 14) than subsampling the random groups (168, SD = 70, SE = 15) and individual PVs were more likely to be resampled from the host-related groups compared with the random groups (Fig. S5 *B* and *D*). Taken together, there is significantly more genetic information relevant to the bovine/human groupings compared with random groups. Second, when the PV selection used to assign the complete human and bovine groups (for Fig. 1*A*) was applied to randomly selected groups, the majority of probabilities were around 0.5 (Fig. S6). Both approaches give us confidence that our capacity to differentiate bovine and cattle isolates is not a result of chance and overfitting from a complex dataset.

US Dataset. We obtained 44 human and 44 bovine isolate sequences from the United States (Dataset S1, US isolates). The isolate distribution based on continental differences is apparent in the phylogenetic tree (Fig. S1). The US isolates occupy separate branches from UK strains, even within the same lineages, and show anticipated bias in host designation with lineage. There were 30 human and 8 bovine strains in lineage I, 13 human and 12 bovine strains in lineage I/II, and 1 human and 23 bovine strains in lineage II. Also, US strains share between them fewer "core" positions, covering only 79% of the Sakai genome. MMDS analysis showed results similar to the UK dataset: The isolates were separated predominantly by lineage (Fig. 2 *A* and *B*). Before testing our UK isolate model across to this dataset, we first built an SVM classifier based only on the US dataset, and the results were similar to UK isolates: The model accuracy was 91.3%, with 92% of the strains assigned correctly according to the host the isolate was from. Four out of 44 bovine isolates were called "human." Thus, even though the US isolates seemed to be distinct in terms of the human/bovine split on the phylogenetic tree and in an MMDS plot, the SVM analysis identified a small group of bovine strains (again just under 10%) that possessed genome features that can be found in the majority of disease-associated human isolates and therefore possibly have greater zoonotic potential. Also, as in the UK dataset, the predicted probabilities of most isolates had little variation, and therefore potentially contain strictly bovine or human features, whereas a smaller group exhibited much greater variability.

In the US dataset, there was a total of 10,590 PVs that varied between the two hosts, which is comparable with the UK dataset (10,878 PVs). However, the US dataset contained a much higher number of PVs with larger differences between hosts (Table S1). However, there was a relatively small overlap of discriminatory PVs (n = 197) between the two datasets. The US dataset was tested with the model trained on the UK dataset based on these 197 PVs. Despite the small number of overlapping PVs between the datasets, the model accuracy was 78%, with 38 out of 44 bovine isolates and 31 out of 44 human isolates assigned according to the host from which they were isolated. When trained on the US dataset and tested on the UK dataset, 86 out of 94 bovine isolates and 78 out of 91 human isolates were assigned to the isolation host. Therefore, even though there are considerable differences between the two datasets and a significant amount of continent-specific information has to be excluded, the same model can be applied, although with less accuracy, to a distinct dataset.

Despite the continental divergence between the UK and US isolates, we tried combining the two datasets for testing. Based on an MMDS analysis, human US isolates that belong to lineage I form a separate cluster far apart from other lineage I isolates (Fig. S7 *A* and *B*); however, the overall tendency is similar for the UK or US datasets alone, with lineage I isolates separated from all of the others. When the proportion of PVs was calculated for the sets combined, some descriptive features from one dataset become neutralized by the other dataset. The SVM model based on the combined dataset (Fig. S7) achieved 82% of model accuracy and predicted 84% of human isolates and 83% of bovine isolates correctly according to their isolation host. It was reassuring that among the bovine isolates that were called "human" were all of these that already were assigned "human" from the single-country models. The same applies to human isolates that were called "bovine." However, the subset of bovine strains called "human" in a mixed model increased potentially due to differences in PVs that define human/bovine separation in the United Kingdom and United States.

UK Outbreaks. Two main hypotheses can be generated from these findings, although they are not mutually exclusive: (*i*) Isolates associated with human infections represent a very specific subset of bovine isolates, in which case the majority of bovine *E. coli* O157 isolates that we have sequenced may be unlikely to cause human disease, and (*ii*) isolates change their genome content following transition into another host, so potentially they acquire phage/plasmid regions in the human host although they originate from cattle; the reverse transfer and adaptation is also possible. To address this question we analyzed EHEC O157 strains from two outbreak investigations. One outbreak was associated with hamburger consumption where both the meat and animal sources were identified (human n = 17, cattle n = 5, and hamburger n = 12). Another was associated with milk consumption (human n = 9 and milk n = 3). As anticipated, the individual outbreak strains closely relate to each other and in the phylogenetic tree formed individual tight clusters for each outbreak (Fig. S1). On an MMDS plot they clustered slightly separately from all other UK strains but in close proximity to lineage I PT 21/28 strains to which they belong (Fig. 3 *A* and *B*).

We trained the model on the all-UK dataset, excluding the outbreak isolates, and tested it on the outbreak isolates. From both outbreaks the "bovine" isolates (from milk, hamburger meat, and cattle) were classified as "human," with probabilities higher than 0.75 for any isolate (Fig. 3*C*). This supports the first hypothesis that the threat to human health originates primarily from a minor subset of strains and that the majority of bovine strains from both our UK and US datasets, despite their core SNP association and virulence gene content, are unlikely to be associated with disease in humans.

Descriptive Proteins. To assess what level of differences can be found at the core SNP versus PV level we selected four "pairs" of isolates that lie in close proximity to each other on the final branches of the phylogenetic tree but were isolated from different hosts and were predicted by the SVM model to be

associated with those hosts. These pairs had from 9 to 26 SNPs between them whereas the number of unique PVs ranged from 137 to 364, and the relative number of unique PVs between the pairs increased in line with the number of SNPs between the pairs (Table S4). This indicated that these PVs were being lost or acquired over relatively recent evolutionary time because the core mutation rate of *E. coli* has been estimated to be two to three SNPs per year (22).

We then summarized the differential PVs across the UK dataset based on their annotations, and for the complete UK dataset with $\Delta PV > 10$ there were 292 PVs that had higher proportions in human compared with bovine isolates (summarized in Fig. 4; "hypothetical proteins" were not included in the figure). By comparison, 343 PVs (20% more) were present in higher proportions in the bovine isolates compared with the human. The main annotated proteins in both groups were similar and were predominately prophage-related proteins. Variation in prophage content therefore underpins the human/bovine classification demonstrated in this study. This accords with expectations about *E. coli* strain evolution being driven by prophage acquisition, rearrangement, and loss. Different prophage annotations do appear depending on the host (i.e., rac prophage with 3% for bovine isolates and dlp12 prophage with 3% for human), although work is now required to examine the biological impact of differential PVs and how these alter the potential of an isolate to infect or cause disease in humans.

Conclusions

This study has applied machine learning to predict the zoonotic potential of bacterial isolates. The analysis demonstrates that in the highly clonal *E. coli* O157 serogroup, host-specific information can be inferred from WGS analysis. Moreover, using an SVM classifier it was possible to generate a probability of host association that indicated that only a minor (<10%) subset of bovine strains were likely to have an impact on human health. In fact, none of the cattle isolates (apart from outbreak trace-back isolates) achieved very high human association probabilities (>0.9), potentially indicating that those posing a serious zoonotic threat are very rare. This finding has implications for public health management of this disease because it means that such strains can now potentially be identified in the ruminant reservoir and, if these are the exception, then targeted control strategies including vaccination or even eradication become a more realistic option to protect human health. The specific prophages that encode the differential PVs now need to be identified to progress our understanding of this zoonosis. A subset of isolates from humans were called as "bovine," and currently we do not know whether they differed in their disease severity, e.g., whether isolates from humans that had high bovine probabilities were more likely to be associated with asymptomatic infections (23). In summary, we consider that machine-learning approaches have tremendous potential to interrogate complex genome information for which specific attributes of the organism, such as disease or isolation host, are known.

Materials and Methods

UK and US datasets were previously studied (UK dataset in ref. 13 and US dataset in ref. 24). Illumina short read sequences were assembled with SPAdes (25) and annotated with Prokka (26). Maximum likelihood (ML) core SNPs trees were constructed with RAxML (27). MMDS was performed as described in ref. 28. Pan-genomes were constructed using Roary (29); the threshold was set to 95% of sequence similarity at the amino-acid level. A classifier based on an SVM algorithm was built using R package e1071 (30). The model was tuned and cost and gamma parameters were adjusted (f.ex to gamma = 1e-04 and cost = 100 for the UK dataset). No review board approval was required for the experiments described in this manuscript. Full details of methods are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We would like to acknowledge the value of human and bovine *E. coli* O157 sequence data available from previous published studies, especially work from the Wellcome Trust IPRAVE consortium, Public Health England, and the Scottish *E. coli* reference laboratory. This work was supported by Food Standards Scotland and the Food Standards Agency Grant FS101055 (to D.L.G., T.J.D., and L.M.), which has allowed the continuation of significant EHEC O157 research in the UK. This research was also supported by a University of Edinburgh studentship (N.L.) and core Biotechnology and Biological Sciences Research Council strategic programme Grant BB/J004227/1 (to D.L.G.). T.J.D. was funded by the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections at the University of Liverpool in partnership with Public Health England, University of East Anglia, University of Oxford, and the Institute of Food Research.

1. Quick J, et al. (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol* 16:114.
2. He M, et al. (2013) Emergence and global spread of epidemic healthcare-associated Clostridium difficile. *Nat Genet* 45(1):109–113.
3. Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. *Curr Opin Microbiol* 23:148–154.
4. Pearce MC, et al. (2006) Prevalence and virulence factors of Escherichia coli serogroups O26, O103, O111, and O145 shed by cattle in Scotland. *Appl Environ Microbiol* 72(1):653–659.
5. Gunzer F, et al. (1992) Molecular detection of sorbitol-fermenting Escherichia coli O157 in patients with hemolytic-uremic syndrome. *J Clin Microbiol* 30(7):1807–1810.
6. Griffin PM, Tauxe RV (1991) The epidemiology of infections caused by Escherichia coli O157:H7, other enterohemorrhagic E. coli, and the associated hemolytic uremic syndrome. *Epidemiol Rev* 13:60–98.
7. Mead PS, Griffin PM (1998) Escherichia coli O157:H7. *Lancet* 352(9135):1207–1212.
8. Wells JG, et al. (1991) Isolation of Escherichia coli serotype O157:H7 and other Shiga-like-toxin-producing E. coli from dairy cattle. *J Clin Microbiol* 29(5):985–989.
9. Borczyk AA, Karmali MA, Lior H, Duncan LM (1987) Bovine reservoir for verotoxin-producing Escherichia coli O157:H7. *Lancet* 1(8524):98.
10. Kim J, Nietfeldt J, Benson AK (1999) Octamer-based genome scanning distinguishes a unique subpopulation of Escherichia coli O157:H7 strains in cattle. *Proc Natl Acad Sci USA* 96(23):13288–13293.
11. Zhang Y, et al. (2007) Genome evolution in major Escherichia coli O157:H7 lineages. *BMC Genomics* 8:121.
12. Manning SD, et al. (2008) Variation in virulence among clades of Escherichia coli O157:H7 associated with disease outbreaks. *Proc Natl Acad Sci USA* 105(12):4868–4873.
13. Dallman TJ, et al. (2015) Applying phylogenomics to understand the emergence of Shiga-toxin-producing Escherichia coli O157:H7 strains causing severe human disease in the UK. *Microbial Genomics*, 10.1099/mgen.0.000029.
14. Pearce MC, et al. (2009) Temporal and spatial patterns of bovine Escherichia coli O157 prevalence and comparison of temporal changes in the patterns of phage types associated with bovine shedding and human E. coli O157 cases in Scotland between 1998-2000 and 2002-2004. *BMC Microbiol* 9:276.
15. Dowd SE, Williams JB (2008) Comparison of Shiga-like toxin II expression between two genetically diverse lineages of Escherichia coli O157:H7. *J Food Prot* 71(8):1673–1678.
16. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20(3):273–297.
17. Yang ZR (2004) Biological applications of support vector machines. *Brief Bioinform* 5(4):328–338.
18. Cruz JA, Wishart DS (2007) Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2:59–77.
19. Hayashi T, et al. (2001) Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* 8(1):11–22.
20. Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA (2011) Genomic anatomy of Escherichia coli O157:H7 outbreaks. *Proc Natl Acad Sci USA* 108(50):20142–20147.
21. Ohnishi M, Kurokawa K, Hayashi T (2001) Diversification of Escherichia coli genomes: Are bacteriophages the major contributors? *Trends Microbiol* 9(10):481–485.
22. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc Natl Acad Sci USA* 109(41):E2774–E2783.
23. Silvestro L, et al. (2004) Asymptomatic carriage of verocytotoxin-producing Escherichia coli O157 in farm workers in Northern Italy. *Epidemiol Infect* 132(5):915–919.
24. Norman KN, Strockbine NA, Bono JL (2012) Association of nucleotide polymorphisms within the O-antigen gene cluster of Escherichia coli O26, O45, O103, O111, O121, and O145 with serogroups and genetic subtypes. *Appl Environ Microbiol* 78(18):6689–6703.
25. Bankevich A, et al. (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comp Biol* 19(5):455–477.
26. Seemann T (2014) Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069.
27. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
28. Abdi H (2007) Metric multidimensional scaling (MDS): Analyzing distance matrices multidimensional scaling: Eigen-analysis of a distance matrix. *Encyclopedia of Measurement and Statistics*, ed Salkind NJ (Sage, Thousand Oaks, CA), Vol 2, pp 598–605.
29. Page AJ, et al. (2015) Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31(22):3691–3693.
30. Meyer D, et al. (2015) Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package e1071, version 1.6-7 (Vienna University of Technology, Vienna).