Pathogenic potential assessment of the Shiga toxin–producing *Escherichia coli* by a source attribution–considered machine learning model

Hanhyeok Im[a,b], Seung-Ho Hwang[a,b], Byoung Sik Kim[c], and Sang Ho Choi[a,b,d,1]

[a]National Research Laboratory of Molecular Microbiology and Toxicology, Seoul National University, 08826 Seoul, Republic of Korea; [b]Department of Agricultural Biotechnology and Center for Food Safety and Toxicology, Seoul National University, 08826 Seoul, Republic of Korea; [c]Department of Food Science and Engineering, Ewha Womans University, 03760 Seoul, Republic of Korea; and [d]Center for Food and Bioconvergence, Seoul National University, 08826 Seoul, Republic of Korea

Edited by Marvin Whiteley, Georgia Institute of Technology, Atlanta, GA, and accepted by Editorial Board Member Caroline S. Harwood March 29, 2021 (received for review September 8, 2020)

Instead of conventional serotyping and virulence gene combination methods, methods have been developed to evaluate the pathogenic potential of newly emerging pathogens. Among them, the machine learning (ML)–based method using whole-genome sequencing (WGS) data are getting attention because of the recent advances in ML algorithms and sequencing technologies. Here, we developed various ML models to predict the pathogenicity of Shiga toxin–producing *Escherichia coli* (STEC) isolates using their WGS data. The input dataset for the ML models was generated using distinct gene repertoires from positive (pathogenic) and negative (nonpathogenic) control groups in which each STEC isolate was designated based on the source attribution, the relative risk potential of the isolation sources. Among the various ML models examined, a model using the support vector machine (SVM) algorithm, the SVM model, discriminated between the two control groups most accurately. The SVM model successfully predicted the pathogenicity of the isolates from the major sources of STEC outbreaks, the isolates with the history of outbreaks, and the isolates that cannot be assessed by conventional methods. Furthermore, the SVM model effectively differentiated the pathogenic potentials of the isolates at a finer resolution. Permutation importance analyses of the input dataset further revealed the genes important for the estimation, proposing the genes potentially essential for the pathogenicity of STEC. Altogether, these results suggest that the SVM model is a more reliable and broadly applicable method to evaluate the pathogenic potential of STEC isolates compared with conventional methods.

STEC | machine learning | risk assessment | pathogenic potential

Emerging pathogens causing an increasing number of outbreaks are now considered as the major risk to public health (1). The exact assessment of the pathogenic potential of pathogens is required to predict and manage their health risk in advance (2). Conventional methods such as serotyping and virulence gene combinations have been used to assess the bacterial pathogenic potentials (3, 4). However, these conventional assessment methods are not reliable for evaluating the pathogenic potential of emerging pathogens, because the same serotype may carry different virulence genes, and/or contribution of unknown virulence genes to the bacterial pathogenicity is still possible (4, 5). Therefore, the development of methods assessing their pathogenic potential is required to cope with the public health risks caused by newly emerging pathogens.

Shiga toxin–producing *Escherichia coli* (STEC) causes a wide range of human illnesses ranging from mild diarrhea to hemolytic uremic syndrome, which often results in permanent kidney failure (6). In addition to the O157 serotype STEC, emerging non-O157 serotype STECs have been identified as causative agents for the increasing outbreaks lately (5, 7). However, the relationships between the non-O157 serotypes and their pathogenicity have not been defined yet, and thus, predicting the pathogenic potential of the non-O157 serotype STECs has limitations (4, 5). It has been reported that virulence genes such as *stx2* and *eae* are required for the pathogenesis of STEC (4, 5, 7–9). However, the emerging highly pathogenic STEC isolates carry novel virulence genes (4, 5), and indeed, a STEC isolate with a novel combination of *stx2* and *aggR* had caused a huge outbreak in Europe in 2011 (7, 10).

Recently, advances in next-generation sequencing technologies have enabled us to exploit whole-genome sequencing (WGS) data (11, 12). Although the WGS data of pathogens can provide rich information about various genetic features of the pathogens, these data are too complex to gain valuable insights on their pathogenicity by using traditional statistical methods (12, 13). In contrast, machine learning (ML) algorithms have notable performance in the analysis of the complex WGS data (12, 13) and therefore have been exploited lately to find out the connection between genetic features and pathogenicity of some pathogens (12, 14–17). The ML algorithms include two broad categories: unsupervised and supervised. The unsupervised ML algorithms, such as phylogenetic tree analysis, principal component analysis (PCA), and Gaussian mixture model (GMM), recognize the inherent patterns in a dataset without the concept of output and then discriminate the given dataset using the inherent patterns (17, 18). On the other hand, the supervised ML algorithms such as Gaussian Naive Bayes

Significance

Outbreaks of newly emerging pathogens have posed serious threats to public health. Here, we developed a machine learning (ML) model to evaluate the pathogenic potential of the Shiga toxin–producing *Escherichia coli* (STEC) isolates using their whole-genome sequencing data. The developed ML model using the support vector machine (SVM) algorithm, the SVM model, correctly estimated the pathogenic potentials of the previous outbreak isolates. The SVM model was also more reliable and widely applicable to predict the pathogenicity of STEC than the conventional assessment methods. Permutation importance analyses further identified the genes potentially associated with the pathogenicity of STEC. This ML-based approach will be a promising assessment method to identify the risk of newly emerging pathogens.

Author contributions: H.I. and S.H.C. designed research; H.I. performed research; H.I. and S.H.C. analyzed data; and H.I., S.-H.H., B.S.K., and S.H.C. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. M.W. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: choish@snu.ac.kr.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2018877118/-/DCSupplemental.

Published May 13, 2021.

MICROBIOLOGY

(GaussianNB), decision trees (DTs), random forest (RF), and support vector machine (SVM) predict an output from an input data. However, these supervised ML algorithms need to be trained on known input–output pairs until they can predict the correct output using the given input data (17).

In this study, we built various ML models and compared their performances of evaluating the pathogenicity of STEC isolates. We subsequently developed an ML model using the SVM algorithm, the SVM model, which can evaluate the pathogenic potential of the STEC isolates the most accurately among the tested ML models. Because the SVM model can also estimate the pathogenic potential of STEC isolates of which the pathogenicity cannot be estimated by conventional methods, the model is more widely applicable to predict the risk of STEC isolates. Moreover, permutation importance analyses discovered the genes important for the evaluation of the SVM model and identified the genes potentially contributing to the pathogenicity of the STEC isolates.

Results

Generation and Validation of the Input Dataset for the ML Models. A large-scale pangenome comprising a total of 22,497 genes was constructed using the WGS data of 2,646 STEC isolates consisting of 2,292 clinical isolates (pathogenic, positive control group) and 354 environmental isolates (nonpathogenic, negative control group), which are classified based on the source attribution, the relative risk potential of the isolation sources (*Materials and Methods*). From the pangenome, the genes statistically relevant to either the positive or negative control group were selected as significant genes by the pangenome-wide association studies (pan-GWAS) (19). As a result, a total of 3,453 significant genes, including 148 virulence genes, were selected (*SI Appendix*, Fig. S1). The 148 virulence genes included the major virulence genes of the pathogenic STEC, such as *eae*, *aggR*, and the locus of enterocyte effacement (LEE) effector protein genes (*SI Appendix*, Table S1) (10, 20, 21). Furthermore, as expected, most of the virulence genes (125/148) were notably involved in the positive control group. These results reflect that the two control groups are indeed classified mainly by the differences in their pathogenic potentials. To further validate the grouping based on the source attribution, the same pan-GWAS was conducted 100 times on the trial groups which were randomly mixed and then divided. As a result, only 285.9 significant genes, including 9.3 virulence genes on average (total 28,592 significant genes, including 933 virulence genes, were divided by 100), were selected (*SI Appendix*, Fig. S1). The reduction of the significant genes indicated that the initial positive and negative control grouping is valid, and the significant genes of the resulting groups are nonaccidental. It has been reported that the subtypes of Shiga toxins are also associated with the pathogenicity of STEC (4, 5). Thus, the 10 Shiga toxin genes, $stx1_a$, $stx1_c$, $stx1_d$, $stx2_a$, $stx2_b$, $stx2_c$, $stx2_d$, $stx2_e$, $stx2_f$, and $stx2_g$, were added to the 3,453 significant genes. Accordingly, the presence/absence matrix of the 3,463 genes of the 2,646 STEC isolates was used as an input dataset of the ML models for further analysis.

The Unsupervised ML Algorithms Cannot Discriminate between the Clinical and Environmental Isolates. To examine whether the ML algorithms can discriminate between the clinical and environmental isolates using the input dataset, the unsupervised ML algorithms were first tested. The phylogenetic tree split the isolates in the input dataset into three clades, which contained the clinical and environmental isolates together (Fig. 1*A*). Although clade I (red box) and clade II (yellow box) mainly grouped the clinical isolates, clade III (blue box) carried a similar ratio of clinical and environmental isolates together (Fig. 1*A*). Consequently, the phylogenetic tree cannot distinguish the clinical and environmental isolates from each other. The PCA plot also revealed several clusters of isolates which were mainly composed of the clinical isolates containing a small number of environmental isolates (Fig. 1*B*). The environmental isolates, however, did not form their own cluster. Most of the environmental isolates were mixed with the clinical isolates and scattered over a broad region (Fig. 1*B*). The models using the GMM algorithm also performed poorly in discriminating the clinical and environmental isolates with a maximum accuracy of 44% (*SI Appendix*, Fig. S2). These

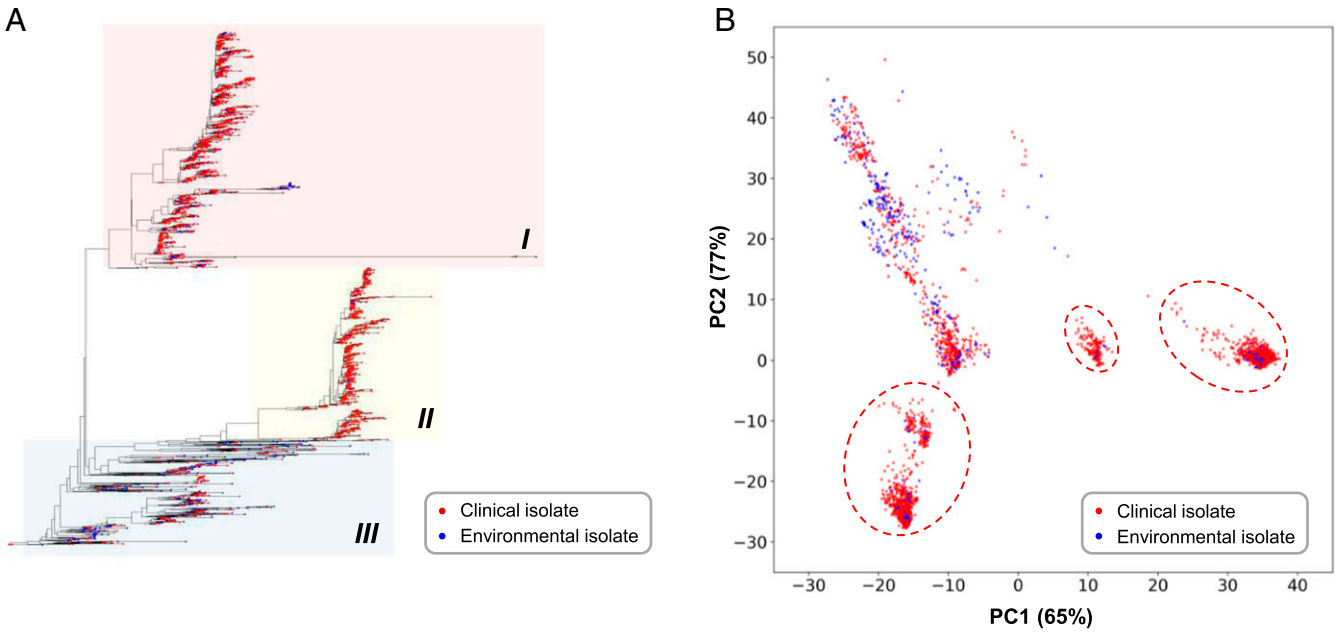

Fig. 1. Analyses of the STEC isolates using the input dataset based on the unsupervised ML algorithms. The red dot represents the clinical isolate, and the blue dot represents the environmental isolate. (*A*) Phylogenetic tree of the STEC isolates based on a maximum likelihood method. The three main clades are emphasized by colored boxes. (*B*) The PCA plot of the STEC isolates. PC1, Principal component 1; PC2, Principal component 2. The clusters primarily comprising the clinical isolates are circled by the dashed red line.

results indicate that the unsupervised ML algorithms cannot effectively discriminate between the clinical and environmental isolates.

The Supervised ML Model Using the SVM Algorithm Most Effectively Discriminates between the Clinical and Environmental Isolates. Four different supervised ML models using the GaussianNB, DTs, RF, and SVM algorithms were trained on each training dataset produced by the stratified 10-fold cross-validation (CV) of the input dataset to discriminate between the clinical and environmental isolates. All the supervised ML models performed on 10 different training and test dataset pairs showed good discrimination performances with accuracy, precision, and true positive rate scores over 0.84 (*SI Appendix*, Fig. S3). These results indicated that the supervised ML models were able to discriminate between the clinical and environmental isolates. The Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUROC) were further exploited to compare the discrimination performances of the supervised ML models. Among them, the SVM model showed the best performance with an MCC score of 0.66 (Fig. 2*A*). The SVM model also presented the steepest receiver operating characteristic (ROC) curve with an area under the curve (AUC) score of 0.93 (Fig. 2*B* and *SI Appendix*, Fig. S4), showing that the SVM model performs best. To confirm that the SVM model performance is valid, the SVM models were trained on the datasets consisting of the significant genes selected from the only training sets produced by the stratified 10-fold CV. The resulting MCC and AUC scores of the SVM models were not different from those of the SVM model trained on the input dataset (*SI Appendix*, Fig. S5 *A* and *B*), demonstrating that the performance of the SVM model is not the result of overfitting to the input dataset. Altogether, these results indicate that the SVM model is the most appropriate supervised ML model to classify the clinical and environmental STEC isolates.

The SVM Model Evaluates the Pathogenic Potential of the STEC Isolates Accurately. Based on the previous assumption that the clinical and environmental isolates represent the pathogenic and nonpathogenic group, respectively, the SVM model calculated the decision function values of each isolate. The isolates with a decision function value either over 0 or under 0 were classified into the pathogenic or nonpathogenic group, respectively. Over 98% of the clinical isolates (positive controls) in the input dataset (2,269/2,292) were classified into the pathogenic group. Similarly, over 96% of the environmental isolates (negative controls) in the input dataset (343/354) were classified into the nonpathogenic group. As shown in Fig. 2*C*, the clinical isolates had decision function values in the first quartile (Q1) 1.00, and the environmental isolates had decision function values in the third quartile (Q3) −1.00, indicating that the distributions of the decision function values were clearly distinguished between the clinical and environmental isolates. The combined results indicate that the SVM model could discriminate between the clinical and environmental isolates correctly and clearly, thereby accurately predicting the pathogenic potential of the STEC isolates using the input dataset.

The SVM Model Evaluates the Pathogenic Potential of the STEC Isolates According to Their Source Attribution and Clinical Outcomes. The environmental isolates from cattle, dairy products, and farm products, the major sources of STEC outbreaks, were previously excluded from the negative control group in the input dataset. The SVM model examined the pathogenic potential of the isolates to prove that their exclusion from the negative control group to construct the input dataset is correct. The cattle isolates showed a broad distribution of decision function values ranging from −1.27 to 2.51 (Fig. 3*A*). Nonetheless, about 80% of the cattle isolates (514/642) had decision function values over 0 and thus were classified into the pathogenic group. Moreover, about 37% of the cattle isolates (235/642) had decision function values even over 1.00, which was comparable with those of the clinical isolates (Fig. 2*C*). The six out of seven dairy product isolates and three out of eight farm product isolates were also classified into the pathogenic group with decision function values over 0 (Fig. 3*A*). The SVM model effectively estimated that many of the environmental isolates from the major sources of the STEC outbreaks are pathogenic as previously reported by the source attribution of STEC (4, 5, 22), suggesting that excluding these isolates from the negative control group is a proper approach to construct the input dataset.

The SVM model was then applied to 83 pathogenic STEC isolates with the history of outbreaks to further validate its assessment results. It should be noted that the outbreak isolates were not included in the input dataset and thus not used in the previous training of the SVM model. Nevertheless, all the outbreak isolates were classified into the pathogenic group by the SVM model, even though the isolates originated from entirely different outbreak cases (Fig. 3*B*). This result indicates that the SVM model correctly evaluates the pathogenic potential of the STEC isolates consistent with their clinical outcomes. Altogether, the combined results suggest that the SVM model is able to

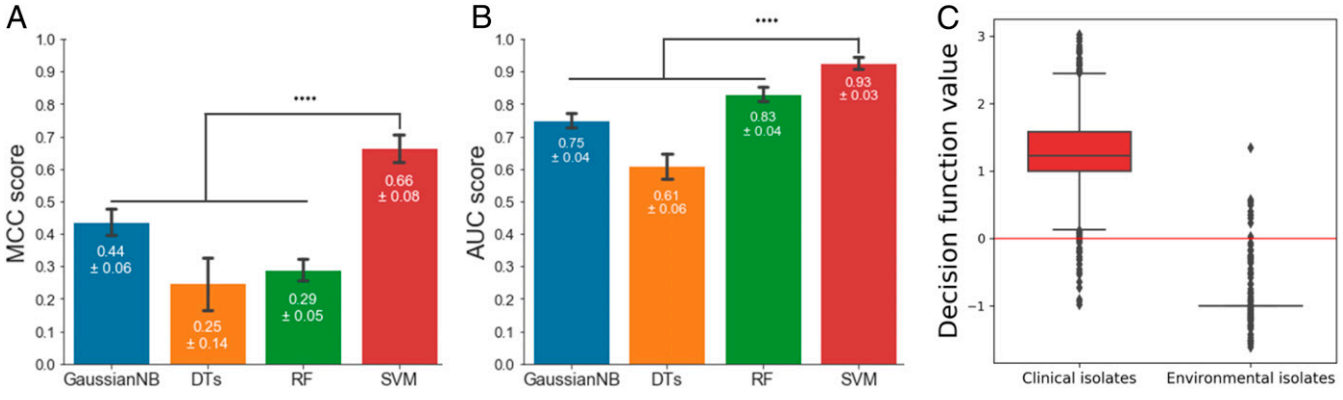

Fig. 2. The discrimination performances of the supervised ML models for the STEC isolates in the input dataset. (*A* and *B*) The bar plots of the discrimination performances of the supervised ML models using four different algorithms: GaussianNB, DTs, RF, and SVM, as indicated. The performances of these models were scored with MCC (*A*) and AUROC (*B*). MCC and AUROC have a score of 1 for a perfect prediction. The average scores of the individual models are indicated at the tip of the bars. SD is represented by the error bar and score. Statistical significance was determined by Student's *t* test (****$P < 0.00005$). (*C*) The box plots of the decision function values of the clinical and environmental isolates in the input dataset calculated by the SVM model. The clinical isolates had decision function values of median 1.22 (Q1, Q3: 1.00, 1.58), and the environmental isolates had decision function values of median −1.00 (Q1, Q3: −0.99, −1.00). The end lines of each box show the Q1 and Q3 of the values.

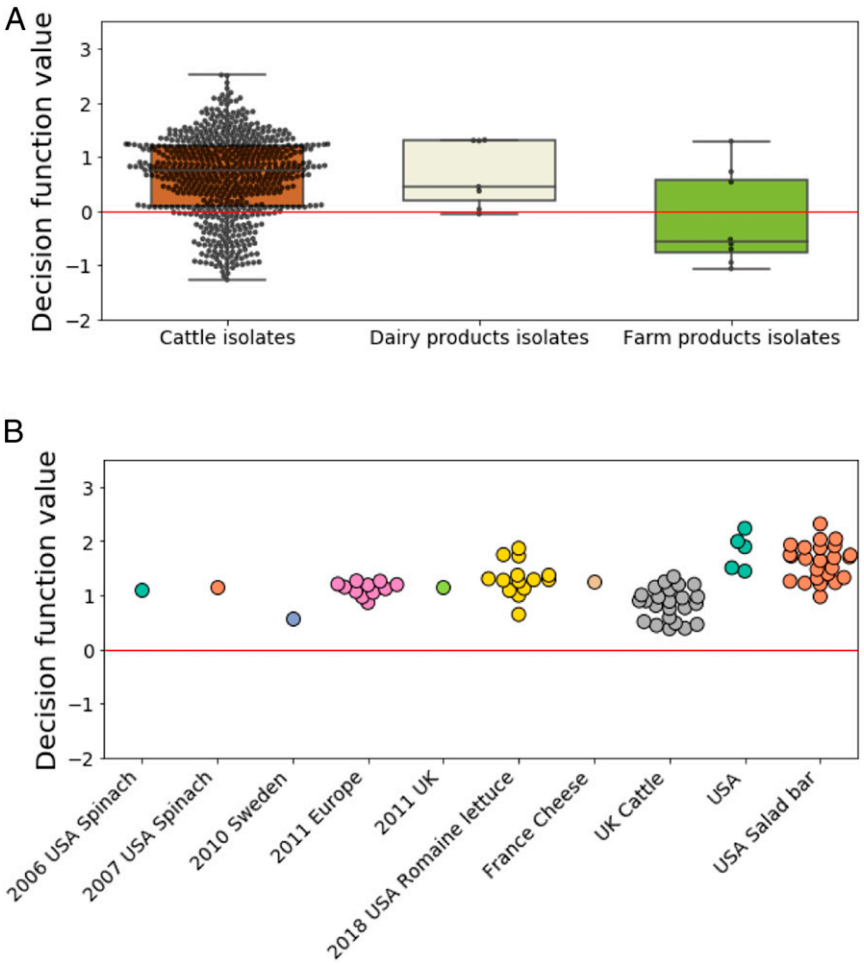

Fig. 3. The box and swarm plots of the decision function values of the isolates associated with the STEC outbreaks. (*A*) The box and swarm plots of the decision function values of the isolates from cattle, dairy products, and farm products that were excluded from the negative control group. Each dot of the plots represents one isolate. The isolates from cattle, dairy products, and farm products had decision function values of median 0.74 (Q1, Q3: 0.10, 1.20), 0.45 (Q1, Q3: 0.20, 1.30), and −0.56 (Q1, Q3: −0.77, 0.58), respectively. The end lines of each box show the Q1 and Q3 of the values. (*B*) The swarm plots of the decision function values of the isolates with the history of outbreaks. The obtainable information about the year, country, and source of the outbreak are labeled as indicated. Each circle of the plots represents one isolate.

produce an effective and reliable assessment of the pathogenic potential of STEC isolates using only their significant gene profiles extracted from the WGS data.

The SVM Model Evaluation Is More Reliable and Broadly Applicable than the Conventional Methods. The isolates in the input dataset were grouped by each serotype and virulence gene combination, and the SVM model estimated the pathogenic potentials of the isolates in each group. The conventional serotyping method showed that the STEC isolates with O26, O157, O121, O145, O111, O104, O91, O103, and O55 serotypes are pathogenic (4, 5, 7, 23, 24). Among the isolates, the SVM model predicted that the isolates of the O26, O157, O121, O145, O111, and O104 serotypes with decision function values of Q1 over 1.00 are pathogenic (Fig. 4*A* and *SI Appendix*, Table S2), comparable with the clinical isolates (Fig. 2*C*). In contrast, the decision function values of Q1 for the isolates with the O91, O103, and O55 serotypes were between 1 and 0 (Fig. 4*A* and *SI Appendix*, Table S2), indicating that the isolates are also pathogenic, but their pathogenicity could be lower than those with decision function values of Q1 over 1. These results revealed that the SVM model can estimate the pathogenicity of the isolates and even can differentiate the pathogenicity with a finer resolution.

Additionally, the SVM model was applied to predict the pathogenic potential of the STEC isolates with the serotypes of which the pathogenicity information is not available. The SVM model estimated that the isolates with the O71, O123, O151, O63, O156, O177, O76, O69, O146, O80, and O182 serotypes had decision function values of Q1 over 0, indicating that most of these isolates are pathogenic (Fig. 4*A* and *SI Appendix*, Table S2). Meanwhile, the isolates with the O128, O84, O45, O21, O5, O113, O165, O136, O22, and O174 serotypes showed decision function values broadly ranging from −1.28 to 2.26 (Fig. 4*B* and *SI Appendix*, Table S2), indicating that these isolates may have varying pathogenic potentials. Most of the isolates with the O8, O75, O130, O139, O109, and O163 serotypes had decision function values under 0 (Fig. 4*B* and *SI Appendix*, Table S2), indicating that these serotype isolates may not be pathogenic. Notably, the isolates that cannot be classified according to their serotypes had high decision function values of Q1 1.00 (Fig. 4*B* and *SI Appendix*, Table S2), indicating that most of these isolates may have high pathogenic potential. Accordingly, the SVM model successfully predicted the pathogenic potential of the STEC isolates, of which the serotype information is not available.

The SVM model then assessed the pathogenic potential of the input dataset isolates carrying distinct virulence gene combinations.

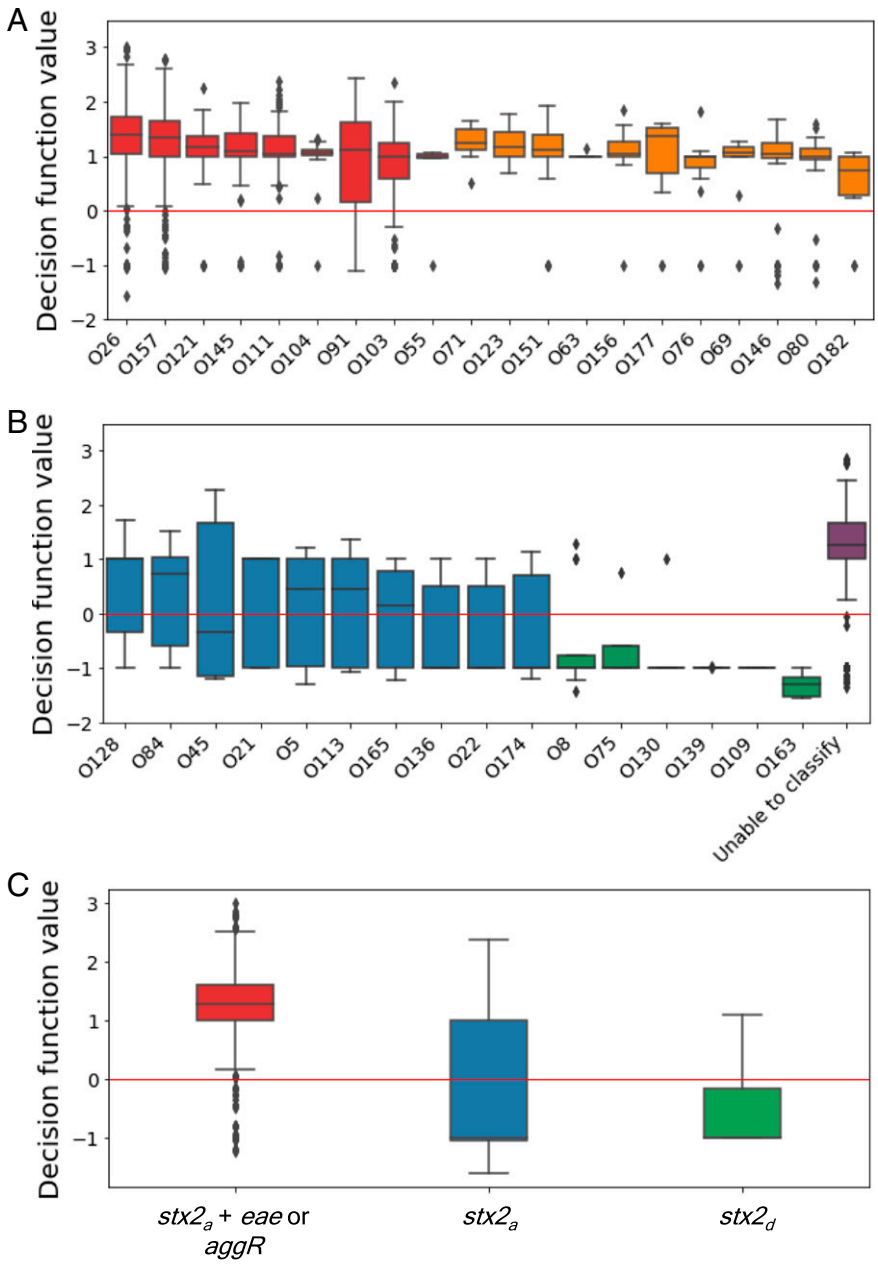

Fig. 4. The box plots of the decision function values of the STEC isolates in the input dataset grouped by the conventional assessment methods. (*A* and *B*) The box plots of the decision function values of the isolates grouped by serotypes. The serotype groups with decision function values of Q1 over 0 (*A*) and the other serotype groups (*B*). The group of isolates that cannot be classified according to their serotype is labeled as "Unable to classify." The serotype groups composed of under five isolates were excluded from the box plots to adjust the figure size. The median, Q1, and Q3 values of the decision function values of the serotype groups can be found in *SI Appendix*, Table S2. (*C*) The box plots of the decision function values of the isolates grouped by virulence gene combinations. The $stx2_a$ + *eae* or *aggR* group, $stx2_a$ group, and $stx2_d$ group had decision function values of a median 1.27 (Q1, Q3: 1.00, 1.61), −1.00 (Q1, Q3: −1.06, 1.00), and −1.00 (Q1, Q3: −1.00, −0.16), respectively. The end lines of each box show the Q1 and Q3 of the values.

The virulence gene combination method showed that the STEC isolates carrying a combination of $stx2_a$ + *eae* or *aggR* are highly pathogenic (4, 5). The SVM model revealed that the isolates with $stx2_a$ + *eae* or *aggR* had decision function values of Q1 over 1.00 and were pathogenic (Fig. 4*C*). The isolates only with $stx2_a$ or $stx2_d$, however, showed a broad spectrum of decision function values ranging from −1.61 to 2.38 (Fig. 4*C*), indicating that these isolates may have varying pathogenic potentials. These results suggest that the virulence gene combinations using Shiga toxin subtypes only have limitations in estimating the pathogenic potential of the STEC isolates. Moreover, about 56% of the isolates in the input dataset (1,504/2,646) do not have such gene combination as $stx2_a$ + *eae* or *aggR*, $stx2_a$, or $stx2_d$, implying that the virulence gene combination method has limited applicability. Consequently, the SVM model is more reliable and broadly applicable than the conventional methods to predict the pathogenic potential of the STEC isolates.

Permutation Importance Analyses Identify the Genes Important to Estimate the Pathogenicity of the STEC Isolates. Permutation importance analysis was conducted for 3,463 input dataset genes and identified 557 genes with the positive weight values, important for

the evaluation of the SVM model performance (Fig. 5*A*). The important genes with the top 25% positive weight values were functionally annotated and primarily assigned to the category of unknown function, followed by the categories of replication, recombination, and transcription (Fig. 5*B*). Only four of the top 20 important genes carry the previously reported functions: antitermination protein Q gene (25, 26), antitermination protein Q homolog gene *quuQ* (25, 26), non-LEE–encoded effector protein gene *espFu* (27, 28), and toxin–antitoxin (TA) system gene *lsoA* (29) (Fig. 5*A* and *SI Appendix*, Table S3). Similarly, permutation importance analysis of 519 input dataset gene clusters identified 182 gene clusters with the positive weight values, revealing their importance. The clusters with the top 25% positive weight values also contained genes that were mostly categorized into the unknown function (Fig. 5 *C* and *D*). The top five important clusters contained a total of 55 genes, including 10 genes with the reported functions: conjugal transfer system *tra* genes (30), sialic acid catabolism *nan* genes (31), TA system genes *phd-doc* (32), and osmotic stress response gene *mscS* (33) (Fig. 5*C* and *SI Appendix*, Table S4). Consequently, the permutation importance analyses of the input dataset identified the genes important to estimate the pathogenicity of the STEC isolates. It is noteworthy that many of the important genes have yet uncharacterized functions, indicating that our SVM model can identify new genes essential for evaluating the pathogenicity of the STEC isolates.

Discussion

To develop the most proper ML model in evaluating the pathogenic potential of the STEC isolates, various ML models were compared by their performances on discriminating between clinical and environmental isolates. In contrast to the unsupervised ML models (Fig. 1 *A* and *B* and *SI Appendix*, Fig. S2), the supervised ML models successfully discriminated between the clinical and environmental isolates using the input dataset (*SI Appendix*, Fig. S3). Among the tested supervised ML models, the SVM model demonstrated the best discrimination performance for the isolates in the test dataset that the model did not previously encounter (Fig. 2 *A* and *B* and *SI Appendix*, Fig. S4). According to the decision function values of the isolates, the SVM model classified most of the clinical and environmental isolates into pathogenic and nonpathogenic groups, respectively (Fig. 2*C*). Additionally, a supervised ML model using the multilayer perceptron (MLP), the MLP model, also effectively discriminated the clinical and environmental isolates, and its accuracies are comparable to those of the SVM model (Fig. 2*C* and *SI Appendix*, Fig. S6*A*). However, the MLP model converged its sigmoid function values to 0 for the nonpathogenic isolates or 1 for the pathogenic isolates (*SI Appendix*, Fig. S6 *A–C*). In contrast, the SVM model calculated the decision function values varying from −1.6 to 3.0 and thus can differentiate the pathogenicity of the isolates (Figs. 2*C* and 3 *A* and *B*). Therefore, the SVM model was more appropriate to estimate the pathogenic potential of STEC isolates with varying degrees using their WGS data only.

An ML model using the WGS data has been developed recently considering isolation host groups and used to predict the isolation hosts of the STEC isolates, assuming that the isolates originating from humans or cattle are pathogenic or nonpathogenic, respectively (14, 15). However, the ML model classified only a minor subset of isolates originating from cattle into the human group as pathogenic and thus might underestimate the pathogenic potential of the cattle isolates, the most common source of STEC outbreaks. Instead, we set the positive and negative control groups by considering the source attribution rather than isolation hosts. Then, the differences in pathogenic potentials present between the two control groups were validated by pan-GWAS (*SI Appendix*, Fig. S1). In contrast to the previous ML model that estimated only under 10% of the cattle isolates to be pathogenic (14, 15), our SVM model estimated about 80% of the cattle isolates to be pathogenic (Fig. 3*A*). In addition, the SVM model also evaluated that many of the isolates from dairy and farm products are pathogenic (Fig. 3*A*). These results also supported that the cattle, dairy products, and farm products could be sources of the pathogenic STEC (4, 5, 22) and thereby should be handled with special care. Moreover, the SVM model correctly predicted the STEC isolates with the history of outbreaks to carry high pathogenic potential (Fig. 3*B*), indicating that the SVM model prediction is

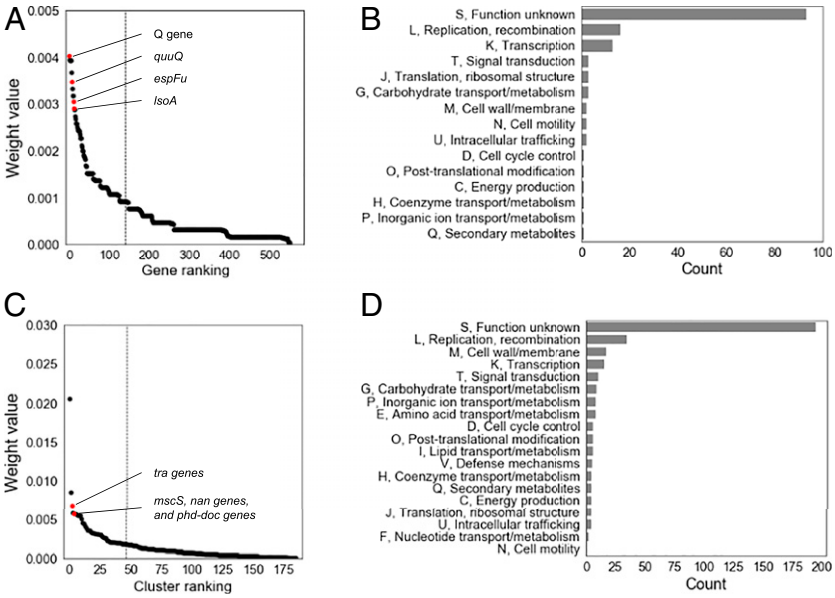

Fig. 5. Permutation importance analyses of the input dataset. (*A* and *C*) The scatter plots of the importance of individual genes (*A*) and correlated gene clusters (*C*). The importance of each gene and cluster is presented by the weight value. The important genes and clusters are plotted by the rank of the positive weight values. The red dots represent the genes with previously reported functions or the clusters, including the genes with the reported functions. The borders of the top 25% important genes and clusters are indicated by dotted lines in the plots. (*B* and *D*) The bar plots of the functional categories of the top 25% of the important genes (*B*) and clusters (*D*). Each category is marked with its alphabetic symbol and functional description.

indeed consistent with the clinical outcome. Accordingly, exploiting the source attribution to establish the positive and negative control groups is a reasonable approach to build an ML model that effectively evaluates the pathogenic potential of STEC isolates.

The SVM model correctly classified the STEC isolates previously designated as pathogenic by the conventional methods into the pathogenic group (Fig. 4 *A* and *C*). In addition, the SVM model further classified the isolates even with the same serotypes or virulence gene combinations into subsets with different decision function values (Fig. 4 *B* and *C*), indicating that the pathogenic potentials of the isolates can be differentiated with a finer resolution using the WGS data. Considering that the isolates with the same serotype predominantly compose a specific clade (23), this result also indicated that the SVM model can even differentiate the pathogenic potentials of the STEC isolates involved in a phylogenetic clade. Moreover, the SVM model could estimate the pathogenic potential of the isolates of which the pathogenicity cannot be evaluated by conventional methods (Fig. 4*B*), revealing its broad applicability. Notably, many of these isolates are predicted to have high pathogenic potential (Fig. 4*B*), emphasizing the necessity of the SVM model rather than conventional methods. The MLP model also correctly classified the STEC isolates previously designated as pathogenic by the conventional methods into the pathogenic group (*SI Appendix*, Fig. S7 *A* and *C*). However, again, the MLP model could not differentiate the pathogenic potential of the isolates with the same serotypes or virulence gene combinations (*SI Appendix*, Fig. S7 *A*–*C*). Consequently, these results suggest that the SVM model using the WGS data are a more precise and applicable method than the conventional methods in evaluating the pathogenic potential of STEC isolates.

The permutation importance analyses identified the genes important for the evaluation of the SVM model. Part of the most important genes with known functions are Q gene, *quuO*, *espFu*, *lsoA*, *phd-doc*, *tra* genes, *nan* genes, and *macS* (Fig. 5 *A* and *C* and *SI Appendix*, Tables S3 and S4). The antitermination protein Q gene and its homolog gene *quuQ* participate in the regulation of Shiga toxin genes (25, 26). The non-LEE–encoded effector protein gene *espFu* is involved in the formation of attaching and effacing lesion, the major mechanism of STEC infection (27, 28). The *lsoA* and *phd-doc* are TA system genes encoded in a plasmid and involved in the anti-phage defense mechanism and maintenance of the plasmid, respectively (29, 32). The *tra* genes are conjugal transfer system genes (30). Considering that the horizontal transfer of plasmids is a major route of STEC to acquire virulence factors (34, 35), these plasmid-encoded and plasmid transfer–related genes possibly contribute to the pathogenicity of STEC. The *nan* genes are the sialic acid catabolism genes (31) and enable the pathogen to utilize the host sialic acids as nutrient sources (31, 36), contributing to the survival and pathogenesis in the host (37). The *mscS* is an osmotic stress response gene (33) and is up-regulated when STEC is exposed to the host intestinal environment (38). Altogether, these results indicate that the SVM model employs the genes associated with the pathogenesis of STEC to estimate its pathogenic potential. However, most of the important genes have yet unknown functions (Fig. 5 *B* and *D*). Nevertheless, new genes significantly associated with the pathogenicity of STEC could be discovered from these important genes with unknown functions, further elucidating the pathogenicity of STEC.

In conclusion, we developed an ML model using the SVM algorithm to effectively estimate the pathogenic potential of STEC isolates using their significant gene profiles extracted from the WGS data, rendering it more extensively applicable than the conventional assessment methods. This study presents an approach exploiting the source attribution to refine the input dataset and thus build an ML model. This ML-based approach could be applied to other pathogens and be used to identify the potential risk of newly emerging pathogens.

Materials and Methods

Generation of the Input Dataset for ML. The WGS data and metadata of 3,303 STEC isolates were retrieved from the GenBank database at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/genbank/) (can be found in Dataset S1 *A* and *B*). The quality of the WGS data were checked using Kraken 2, a taxonomy classification tool (39), and QUAST, a quality assessment tool for genome assemblies (40). The quality-passed WGS data were annotated using Prokka, a prokaryotic genome annotation program (41). The classification of the clinical and environmental isolates was determined based on the metadata of the isolates. The 2,292 clinical STEC isolates were set as the positive control group (pathogenic). The STEC isolates from the cattle, dairy products, and farm products have been reported as the major sources of outbreaks, which may have high pathogenic potential (4, 5, 22). Therefore, among the 1,011 environmental isolates, 657 isolates from major sources of outbreaks were excluded according to the source attribution, and then the remaining 354 environmental isolates were set as the negative control group (nonpathogenic). The pangenome of the STEC isolates was constructed using Pangenome Iterative Refinement and Threshold Evaluation (PIRATE), a pangenomics toolbox (42). The genes statistically relevant to either positive or negative control group were selected as significant genes by Scoary, a pan-GWAS tool ($P < 0.05$) (19) and used to generate the input dataset. The pan-GWAS results were visualized as bar plots with the Seaborn Python packages (https://seaborn.pydata.org/). The significant genes were accurately reannotated using the reference sequences of the UniProt Knowledgebase (UniProtKB) and the Virulence Factor Database (VFDB) (43, 44). The subtypes of the Shiga toxin were identified using the reference sequences of the Shiga toxin subtypes (45). Sequence alignments with the reference sequences were conducted by using Double Index Alignment of Next-generation sequencing Data (DIAMOND), a sequence alignment tool (46). The presence/absence matrix of the significant genes and Shiga toxin subtype genes was used as the input dataset and can be found in Dataset S2.

Unsupervised ML: Phylogenetic Tree, PCA, and GMM. The phylogenetic tree of the input dataset was generated based on the maximum likelihood method with Randomized Axelerated Maximum Likelihood (RAxML), a tool for phylogenetic analysis (47). The reliability of internal branches was assessed by bootstrapping based on 500 replicates. The phylogenetic tree was visualized by the ggtree R package (48). PCA for the input dataset was conducted with the Scikit-learn Python package (49). The GMM models were built using the Scikit-learn Python Package (49) and trained on the PCA-transformed input dataset. The Matplolib Python package was used to visualize the PCA and GMM model results (50).

Supervised ML: GaussianNB, DTs, RF, and SVM. Four different supervised ML algorithms, GaussianNB (51), DTs (52), RF (53), and SVM (54), were used. To select the most appropriate algorithms, the input dataset was randomly split into 90% for training and 10% for test datasets 10 times by stratified sampling, generating 10 different training and test dataset pairs. The optimal hyperparameters were selected via a grid search and used to build the optimized supervised ML models (can be found in *SI Appendix, Materials and Methods*). The optimized supervised ML models were trained on each training dataset by stratified 10-fold CV, and then their discrimination performances were examined using each test dataset. The accuracy, precision, and true positive rate scoring methods were used to compare the performances of the supervised ML models. Because our input dataset consisting of 2,292 clinical isolates and 354 environmental isolates was imbalanced, the MCC and AUROC scoring methods were further used to produce a more informative and truthful score (55, 56). AUROC demonstrates the performance of a certain model as a ROC curve and AUC score. When the ROC curve of a model is steeper, the AUC score is larger, indicating that the model performs better. All of the model buildings, grid searches, and score calculations were conducted with Scikit-learn Python package (49). The bar plots of the scores and the ROC curve plots were visualized by the Seaborn Python packages (https://seaborn.pydata.org/).

Examination of the SVM Model Using the Decision Function Values. The SVM model calculates decision function values to classify each STEC isolate into either pathogenic or nonpathogenic group. If the decision function value was over 0 or under 0, the isolates were classified into the pathogenic or nonpathogenic group, respectively. A greater absolute decision function value indicates a higher confidence score for the classification of an isolate (57). The decision function values of the clinical and environmental isolates in the input dataset were plotted as box plots. The decision function values of the isolates from cattle, dairy products, and farm products were plotted as box and swarm plots.

The WGS data of the 83 isolates with a history of outbreaks were obtained from the BioProject database at the NCBI (https://www.ncbi.nlm.nih.gov/bioproject/) (can be found in Dataset S3), and their decision function values were plotted as swarm plots. The serotypes of the STEC isolates were identified with the SerotypeFinder Center for Genomic Epidemiology (CGE) tool (58). The conventional virulence gene combination method used the combinations of *stx2* subtype a ($stx2_a$) or d ($stx2_d$) with an additional adherence factor *eae* or *aggR* (4, 5). Thus, the isolates were grouped by the following combinations: $stx2_a$, $stx2_a$ + *eae* or *aggR*, and $stx2_d$ ($stx2_d$ + *eae* or *aggR* combination did not exist). The decision function values of the isolates in each serotype and virulence gene combination group were plotted as box plots. All of the plots were visualized by the Seaborn Python packages (https://seaborn.pydata.org/).

Permutation Importance Analyses of the Input Dataset. Permutation importance analysis calculates the importance of a gene of an input dataset by measuring the decrease of the model performance when the data of the gene are shuffled and thus become insignificant (59). The analysis, however, tends to underestimate the importance of the genes which are highly correlated to others (60). Thus, we generated gene clusters based on the Spearman rank–order correlation and used the input dataset gene clusters to figure out the importance of the correlated genes. The permutation importance analyses were repeated 10 times for each gene or cluster and scored their importance, a decrease of the MCC score, as a weight value. The gene clustering and permutation importance analysis were performed with the Scikit-learn Python package (49). Functional categories of the genes were assigned based on the clusters of orthologous group proteins database with the eggNOG-mapper, a functional annotation tool (61). The results of the permutation importance and functional annotation analysis were visualized as scatter plots and bar plots using the Seaborn Python packages (https://seaborn.pydata.org/). The weight values and the functional categories of the input dataset genes can be found in Dataset S4 *A* and *B*.

Data Availability. The data and code used for analysis are available in the GitHub repository at https://github.com/hanhyeok/STEC_pathogenicity_prediction. All other study data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. This work was supported by the National Research Foundation of Korea and funded by the Ministry of Science, Information and Communications Technology (ICT) and Future Planning (2017R1E1A1A01074639).

1. M. Vouga, G. Greub, Emerging bacterial pathogens: The past and beyond. *Clin. Microbiol. Infect.* **22**, 12–21 (2016).
2. J. L. Smith, P. M. Fratamico, Emerging and Re-emerging foodborne pathogens. *Foodborne Pathog. Dis.* **15**, 737–757 (2018).
3. R. V. Tauxe, Emerging foodborne pathogens. *Int. J. Food Microbiol.* **78**, 31–41 (2002).
4. K. Koutsoumanis *et al*., Pathogenicity assessment of Shiga toxin-producing *Escherichia coli* (STEC) and the public health risk posed by contamination of food with STEC. *EFSA J.* **18**, 1–105 (2020).
5. World Health Organization, *Shiga Toxin-Producing Escherichia coli (STEC) and Food: Attribution, Characterization, and Monitoring: Report* (World Health Organization, 2018).
6. M. A. Karmali, Emerging public health challenges of Shiga toxin-producing *Escherichia coli* related to changes in the pathogen, the population, and the environment. *Clin. Infect. Dis.* **64**, 371–376 (2017).
7. European Food Safety Authority, Scientific Opinion on VTEC-seropathotype and scientific criteria regarding pathogenicity assessment. *EFSA J.* **11**, 3138 (2013).
8. P. M. Fratamico *et al*., Advances in molecular serotyping and subtyping of *Escherichia coli*. *Front. Microbiol.* **7**, 644 (2016).
9. U. Naseer, I. Løbersli, M. Hindrum, T. Bruvik, L. T. Brandal, Virulence factors of Shiga toxin-producing *Escherichia coli* and the risk of developing haemolytic uraemic syndrome in Norway, 1992-2013. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 1613–1620 (2017).
10. N. Boisen, A. R. Melton-Celsa, F. Scheutz, A. D. O'Brien, J. P. Nataro, Shiga toxin 2a and Enteroaggregative *Escherichia coli*–A deadly combination. *Gut Microbes* **6**, 272–278 (2015).
11. E. Franz *et al*., Exploiting the explosion of information associated with whole genome sequencing to tackle Shiga toxin-producing *Escherichia coli* (STEC) in global food production systems. *Int. J. Food Microbiol.* **187**, 57–72 (2014).
12. D. Houle, D. R. Govindaraju, S. Omholt, Phenomics: The next challenge. *Nat. Rev. Genet.* **11**, 855–866 (2010).
13. X.-S. Yang, S. Lee, S. Lee, N. Theera-Umpon, Information analysis of high-dimensional data and applications. *Math. Probl. Eng.* **2015**, 1–2 (2015).
14. N. Lupolova, T. J. Dallman, L. Matthews, J. L. Bono, D. L. Gally, Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 11312–11317 (2016).
15. N. Lupolova, T. J. Dallman, N. J. Holden, D. L. Gally, Patchy promiscuity: Machine learning applied to predict the host specificity of *Salmonella enterica* and *Escherichia coli*. *Microb. Genom.* **3**, e000135 (2017).
16. D. Moradigaravand *et al*., Prediction of antibiotic resistance in *Escherichia coli* from large-scale pan-genome data. *PLoS Comput. Biol.* **14**, e1006258 (2018).
17. N. Lupolova, S. J. Lycett, D. L. Gally, A guide to machine learning for bacterial host attribution using genome sequence data. *Microb. Genomics* **5** e000317 (2019).
18. C. M. Svensson, S. Krusekopf, J. Lücke, M. Thilo Figge, Automated detection of circulating tumor cells with naive Bayesian classifiers. *Cytometry A* **85**, 501–511 (2014).
19. O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).
20. J. B. Kaper, J. P. Nataro, H. L. T. Mobley, Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2**, 123–140 (2004).
21. A. R. Pacheco, V. Sperandio, Shiga toxin in enterohemorrhagic *E.coli*: Regulation and novel anti-virulence strategies. *Front. Cell. Infect. Microbiol.* **2**, 81 (2012).
22. World Health Organization, *Enterohaemorrhagic Escherichia coli in Raw Beef and Beef Products: Approaches for the Provision of Scientific Advice: Meeting Report* (World Health Organization, 2011).
23. I. Eichhorn *et al*., Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl. Environ. Microbiol.* **81**, 7041–7047 (2015).
24. L. H. Gould *et al*.; Emerging Infections Program Foodnet Working Group, Increased recognition of non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States during 2000-2010: Epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog. Dis.* **10**, 453–460 (2013).
25. H. Brüssow, C. Canchaya, W.-D. Hardt, Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).
26. S. R. Steyert *et al*., Comparative genomics and stx phage characterization of LEE-negative Shiga toxin-producing *Escherichia coli*. *Front. Cell. Infect. Microbiol.* **2**, 133 (2012).
27. F. H. Martins *et al*., EspFu-mediated actin assembly enhances enteropathogenic *Escherichia coli* adherence and activates host cell inflammatory signaling pathways. *mBio* **11**, 1–18 (2020).
28. F. H. Martins, R. Nepomuceno, R. M. F. Piazza, W. P. Elias, Phylogenetic distribution of tir-cytoskeleton coupling protein (*tccP* and *tccP2*) genes in atypical enteropathogenic *Escherichia coli*. *FEMS Microbiol. Lett.* **364**, 1–7 (2017).
29. Y. Otsuka, T. Yonesaki, Dmd of bacteriophage T4 functions as an antitoxin against *Escherichia coli* LsoA and RnlA toxins. *Mol. Microbiol.* **83**, 669–681 (2012).
30. M. Zatyka, C. M. Thomas, Control of genes for conjugative transfer of plasmids and other mobile elements. *FEMS Microbiol. Rev.* **21**, 291–319 (1998).
31. K. A. Kalivoda, S. M. Steenbergen, E. R. Vimr, Control of the *Escherichia coli* sialoregulon by transcriptional repressor NanR. *J. Bacteriol.* **195**, 4689–4701 (2013).
32. H. Lehnherr, E. Maguin, S. Jafri, M. B. Yarmolinsky, Plasmid addiction genes of bacteriophage P1: *Doc*, which causes cell death on curing of prophage, and *phd*, which prevents host death when prophage is retained. *J. Mol. Biol.* **233**, 414–428 (1993).
33. E. Bremer, R. Krämer, Responses of microorganisms to osmotic stress. *Annu. Rev. Microbiol.* **73**, 313–334 (2019).
34. A. Caprioli, S. Morabito, H. Brugère, E. Oswald, Enterohaemorrhagic *Escherichia coli*: Emerging issues on virulence and modes of transmission. *Vet. Res.* **36**, 289–311 (2005).
35. Y. Ogura *et al*., Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic *Escherichia coli* strains of O157 and non-O157 serotypes. *Genome Biol.* **8**, R138 (2007).
36. K. A. Kalivoda, S. M. Steenbergen, E. R. Vimr, J. Plumbridge, Regulation of sialic acid catabolism by the DNA binding protein NanR in *Escherichia coli*. *J. Bacteriol.* **185**, 4806–4815 (2003).
37. E. R. Vimr, Unified theory of bacterial sialometabolism: How and why bacteria metabolize host sialic acids. *ISRN Microbiol.* **2013**, 816713 (2013).
38. R. Pieper *et al*., Proteomic view of interactions of Shiga toxin-producing *Escherichia coli* with the intestinal environment in gnotobiotic piglets. *PLoS One* **8**, e66462 (2013).
39. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
40. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
41. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
42. S. C. Bayliss, H. A. Thorpe, N. M. Coyle, S. K. Sheppard, E. J. Feil, PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* **8**, 1–9 (2019).
43. UniProt Consortium, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
44. B. Liu, D. Zheng, Q. Jin, L. Chen, J. Yang, VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **47**, D687–D692 (2019).
45. F. Scheutz *et al*., Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J. Clin. Microbiol.* **50**, 2951–2963 (2012).
46. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
47. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

48. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T. Lam, GGTREE: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
49. G. Varoquaux *et al*., Scikit-learn. *GetMobile Mob. Comput. Commun.* **19**, 29–33 (2015).
50. J. D. Hunter, Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
51. A. H. Jahromi, M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features" in *2017 Artificial Intelligence and Signal Processing Conference (AISP)* (IEEE, Piscataway, NJ, 2017), pp. 209–212.
52. S. B. Kotsiantis, Decision trees: A recent overview. *Artif. Intell. Rev.* **39**, 261–283 (2013).
53. M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114**, 24–31 (2016).
54. H. Bhavsar, M. H. Panchal, A review on support vector machine for data classification. *Int. J. Adv. Res. Comput. Eng. Technol.* **1**, 2278–1323 (2012).
55. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
56. D. Berrar, P. Flach, Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief. Bioinform.* **13**, 83–97 (2012).
57. J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **10**, 61–74 (1999).
58. K. G. Joensen, A. M. M. Tetzschner, A. Iguchi, F. M. Aarestrup, F. Scheutz, Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 2410–2426 (2015).
59. A. Altmann, L. Toloşi, O. Sander, T. Lengauer, Permutation importance: A corrected feature importance measure. *Bioinformatics* **26**, 1340–1347 (2010).
60. L. Toloşi, T. Lengauer, Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics* **27**, 1986–1994 (2011).
61. J. Huerta-Cepas *et al*., Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).