Contents lists available at ScienceDirect

International Journal of Food Microbiology

journal homepage: www.elsevier.com/locate/ijfoodmicro

# Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic *Escherichia coli*

Patrick Murigu Kamau Njage*, Pimlapas Leekitcharoenphon, Tine Hald

*Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kemitorvet, Building 204, 2800 Kgs. Lyngby, Denmark*

ARTICLE INFO

*Keywords:*
Hazard characterization
Hazard identification
Risk characterization
Infection outcome
STEC
Whole genome sequencing
Logit boost

ABSTRACT

The ever decreasing cost and increase in throughput of next generation sequencing (NGS) techniques have resulted in a rapid increase in availability of NGS data. Such data have the potential for rapid, reproducible and highly discriminative characterization of pathogens. This provides an opportunity in microbial risk assessment to account for variations in survivability and virulence among strains. A major challenge towards such attempts remains the highly dimensional nature of genomic data versus the number of isolates. Machine learning-based (ML) predictive risk modelling provides a solution to this “curse of dimensionality” while accounting for individual effects that are dependent on interactions with other genetic and environmental factors. This pilot study explores the potential of ML in the prediction of health endpoints resulting from shigatoxigenic *E. coli* (STEC) infection.

Accessory genes in amino acid sequences were used as model input to predict and differentiate health outcomes in STEC infections including diarrhea, bloody diarrhea, hemolytic uremic syndrome and their combinations. Outcomes severity was also distinguished by hospitalization. A matrix of percent similarity between accessory genes and the *E. coli* genomes was generated and subsequently used as input for ML. The performances of ML algorithms random forest, support vector machine (radial and linear kernel), gradient boosting, and logit boost were compared. Logit boost was the best model showing an outcome prediction accuracy of 0.75 (95% CI: 0.60, 0.86), an excellent or substantial performance (Kappa = 0.72). Important genetic predictors of riskier STEC clinical outcomes included proteins involved in initial attachment to the host cell, persistence of plasmids or genomic islands, conjugative plasmid transfer and formation of sex pili, regulation of locus of enterocyte effacement expression, post-translational acetylation of proteins, facilitation of the rearrangement or deletion of sections within the pathogenic islands and transport macromolecules across the cell envelope. We propose further studies are proposed on the proteins with undefined or unclear functionality. One protein family in particular predicted HUS outcome. Toxin-antitoxin systems are potential stress adaptation markers which may mediate environmental persistence of strains in diverse sources.

We foresee the application of ML approach to the set-up of real-time online analysis of whole genome sequence data to estimate the human health risk at the population or strain level. The ML approach is envisaged to support the prediction of more specific STEC clinical endpoints type by inputting isolate sequence data.

## 1. Introduction

Shiga toxin-producing *E. coli* (STEC) represent a diverse category of enteric pathogens associated with gastrointestinal disease of varying severity such as diarrhea, severe diarrhea (hemorrhagic colitis), hemolytic uremic syndrome (HUS), other chronic sequelae following infection such as irritable bowel syndrome and end-stage renal disease or death (Franz et al., 2014; Spinale et al., 2013).

The population structure and emergence of STEC which are not part of the common O157 STEC are increasingly recognized and this has complicated the realization of accurate microbial risk assessment (MRA) (Franz et al., 2014). For successful infection and illness by STEC, the sequence of events include ingestion, survival through the acidic upper gastrointestinal (GI) tract, and colonization of the lower GI tract (Thorpe, 2004). The locus of enterocyte effacement pathogenicity island (LEE locus) has been described as a mechanism by which the lower

[*] Corresponding author.
*E-mail address:* panj@food.dtu.dk (P.M.K. Njage).

https://doi.org/10.1016/j.ijfoodmicro.2018.11.016
Received 28 February 2018; Received in revised form 23 October 2018; Accepted 17 November 2018
Available online 04 December 2018

GI tract is colonized (Paton and Paton, 1998). LEE positive STEC have therefore been associated with the majority of STEC outbreaks. However, a diverse LEE negative non-O157 STEC have been implicated in severe illness outbreaks in Europe and the United States (Buvens et al., 2012; Cooper et al., 2014; Gould et al., 2013; Johnson et al., 2006; Preußel et al., 2013). The variation in disease risk from LEE-positive STEC coupled with the association of some LEE-negative STEC strains with illness suggests the possible role of "exchangeable effector loci" in STEC pathogenesis (Coombes et al., 2008; Tobe et al., 2006). It is therefore important for microbial risk assessment to account for variations in risk of illness by STEC due to population structure and emergence of new STEC strains.

Seropathotype classification has been proposed in efforts to account for the diversity of disease risk outcomes ranging from asymptomatic infection, mild diarrhea, to severe disease such as HUS and hemorrhagic colitis (Karmali et al., 2003). However broad grouping into serotypes does not account for variations in epidemiological outcomes such as disease risk or geographic variation in the distribution of strains. Molecular sub-typing tools provide an opportunity to refine risk assessment by taking into account detailed variations in strains and the associated disease risk between strains within STEC. Molecular biology methods such as microarray and high-throughput PCR systems targeting virulence profiles exist (Bruant et al., 2006; Bugarel et al., 2010; Gonzales et al., 2011). Moreover, whole genome sequencing provides a greater potential input for improved microbial risk assessment because data are not restricted to a specific choice of target genes. Sequence data can also yield further details such as serotype, virulence and antibiotic resistance profiles, and genetic variations such as SNPs. Furthermore, genomic sequences coupled with epidemiological outcomes provide a potential for the discovery of additional biomarkers explaining variation in risk for pathogenic *E. coli*. This provides an opportunity to harness the recent increase in throughput and decreased cost of whole genome sequencing, which has resulted in a rapid increase in available sequence data (Leekitcharoenphon et al., 2014; Pielaat et al., 2013).

Use of whole genome sequencing data for risk assessment in STEC has the potential to support more effective risk-based monitoring protocols, effective public and veterinary health actions and clinical management (Franz et al., 2014). WGS provides an opportunity in MRA for rethinking the classical MRA steps namely hazard identification, exposure assessment, and hazard characterization (Brul et al., 2012).

Several authors have recently reviewed the steps towards use of WGS data in MRA (Brul et al., 2012; Carriço et al., 2013; Havelaar et al., 2010; Pielaat et al., 2013) and an initial example has been reported (Pielaat et al., 2015). The defining step towards the use of WGS data in MRA involves deriving the association between WGS data or its derivatives to health outcomes. However, the high dimensionality issue posed by the high number of potential predictors from WGS data in relation to the number of isolates or sample size presents a key challenge in the application of statistical modelling techniques. Statistical models are either poorly fitting or over-fitting in such circumstances whereas attempts to use data reduction methods may result in biologically less meaningful inference or the loss of important predictors (Houle et al., 2010). Network analysis techniques (Okser et al., 2013) and machine-learning algorithms (Breiman, 2001; Bureau et al., 2005; Houle et al., 2010) have recently provided a family of techniques to model highly dimensional datasets for risk prediction and to derive features (e.g. genes) important for these predictions.

Machine learning algorithms are computer algorithms that improve with experience (Libbrecht and Noble, 2015). These algorithms are potentially robust methods for risk prediction with respect to microbial pathogenesis which is often driven by genetically complex microbial variations and their interactions. Such interactions are averaged out by the use of statistical association methods (Okser et al., 2013). Risk assessment based on machine learning algorithms enable the consideration of both individual predictors as well as interactions with other predictors, which may appear less relevant but nevertheless important in unraveling the strain diversity and the associated variation in phenotypic outcomes (Okser et al., 2013). MRA based on such algorithms provides an opportunity to capture genetic variations acquired over time, thereby, contributing to the early identification of strains with new virulence characteristics. Machine learning algorithms have been instrumental in cancer research progress, where important information has been revealed including patient genotypes, gene expression related phenotypes and patient outcomes (Griffith et al., 2013; Libbrecht and Noble, 2015; Shipp et al., 2002; Whitney et al., 2015). The use of machine learning algorithms in exploring genetic determinants of antimicrobial resistance has been previously reported (Davis et al., 2016; Drouin et al., 2014; Santerre et al., 2015).

We describe in this pilot study a hazard characterization approach in support of increased precision risk assessment applying WGS data for strain rather than whole taxon specific hazard identification, exposure assessment and future illness outcome specific dose-response relationships in risk assessments. Further inference is made concerning relevant genetic features from a complex WGS data that uniquely define STEC clinical outcomes and the survival potential of isolates in differing environments. We envisage this as a first step towards the set-up of web-based tools for the analysis of WGS data from STEC with an aim of predicting the epidemiological risk or health burden at the strain level. The ML approach described here is foreseen as a more specific hazard characterization tool enabling the prediction of the STEC clinical endpoints including diarrhea, bloody diarrhea and HUS or their combinations given isolate sequence data.

## 2. Materials and methods

### 2.1. STEC strains

STEC isolates used in this study were collected over 5 years from a project recently reported by Holmes et al. (2015). This strain panel included: (i) 10 cases collected over 11 months from an outbreak linked to unwashed vegetables in the United Kingdom (UK), (ii) isolates with place and time epidemiological association (8 clusters from single-households, 1 cluster from two farm related households, and 1 cluster that was linked to travel), and (iii) sporadic isolates from 27 patients which were possibly related to travel outside the UK. The metadata associated with the study, DNA isolation, sequencing and availability of the sequencing data are outlined in the manuscript and in the Supplemental material by Holmes et al. (2015) available at http://jcm.asm.org/content/53/11/3565/suppl/DCSupplemental. The diverse strains belonged to ten different multilocus sequence types (STs). A bottleneck in the use of WGS data for MRA is the lack of reproducible health endpoints linking genotypic to phenotypic data. This dataset was unique in that patients had been interviewed for information including severity of infection, travel history, epidemiological linkage and specific source of infection (Holmes et al., 2015). It was, therefore, possible to use clinical endpoints including diarrhea, bloody diarrhea and HUS or their combinations as model outcome or dependent variables in the risk assessment. The severity of these outcomes can also be distinguished by cases requiring hospitalization (Holmes et al., 2015; Preußel et al., 2013) and the clinical outcomes were therefore further sub-categorized into clinical outcome accompanied by hospitalisation versus not accompanied by hospitalization.

### 2.2. Bioinformatics

The model input consisted of accessory genes in form of amino acid sequences. A previously reported approach was used for identification of gene clusters (Binnewies et al., 2005; Friis et al., 2010). In order to obtain gene families, predicted genes in amino acid sequences were determined based on the assembled genomes of the *E. coli* dataset using Prodigal (version 2.5.0) which is a software for gene recognition and translation initiation site identification in prokaryotic genomes (Hyatt et al., 2010). Predicted genes were aligned all-against-all using BLASTP, a Basic Local Alignment search tool (NCBI-blast version 2.2.31 + ) (Camacho et al., 2009) using protein sequences whose query and database sequences are protein sequences. Genes were grouped into the same gene family, if the alignment length and percent similarity were at least 50% (the "50/50 rule"). A blast hit (alignment hit) was considered significant if the alignment covered at least 50% of both sequences, and contained at least 50% identities. Gene families from all genomes were compared. Core genes were built from the intersection of gene families shared by every genome in the analysis. Any gene family that was not part of the core genes was considered as an accessory gene. The number of the core and accessory genes was 2739 and 5737 genes respectively. The size of accessory genes was 2.9 MB. A matrix of percent similarity between the 5737 accessory genes in amino acid sequences and the *E. coli* genomes was generated and subsequently used as input for further modelling.

### 2.3. Predictive modelling

The aim was to link genetic composition of the STEC strains with clinical outcome in humans. Machine learning algorithms were applied as a design-learn-test protocol (Libbrecht and Noble, 2015). We hypothesize that the machine learning models can recognize certain genetic patterns from the input data and further use this to predict outcomes in an unknown sample where only sequence data are presented. Fig. 1 presents an illustration of the machine learning workflow.

Machine learning models include supervised, semi-supervised or unsupervised learning algorithms. We used supervised learning which allows the classification of patterns in the dataset (also referred to as instances or features) into a set of categories (also referred to as *classes* or *labels*) (Rokach, 2010). Classification models (also known as classifiers) are induced from a set of pre-classified patterns (Rokach, 2010). A *training set* which is a subset of the original instances whose labels are known is used to construct an algorithm (inducer) whose particular instance is referred to as a classifier. This classifier labels new instances in the dataset after learning from the known instances in the training set.

Classification algorithms were used for the discrete categories of illness outcome. Classification of new categories involves the identification of the discrete class of a new observation from a training set. Classification is made using the decision $y_c = f_c(X, \theta_c)$, $y_c \epsilon Z$ where $X$ is the new observation's feature vector, $y_c$ is the new observation's category, $f_c(.)$ is the classification function resulting from the training, $\theta_c$ is the parameter set for $f_c(.)$ and Ƶ is the set of class labels (Ren et al., 2016). For instance the aim may be to build a classification system for *E. coli* isolates in this study which may be associated with clinical endpoint HUS ($y_{HUS}$) from the set of clinical endpoints (Ƶ) consisting of diarrhea, bloody diarrhea and HUS or their combinations. This classification function $f_c(.)$ can be used to predict the class of an unknown isolate which in this case is illness outcomes following STEC infection.

Modelling was performed using ensemble methods. Among

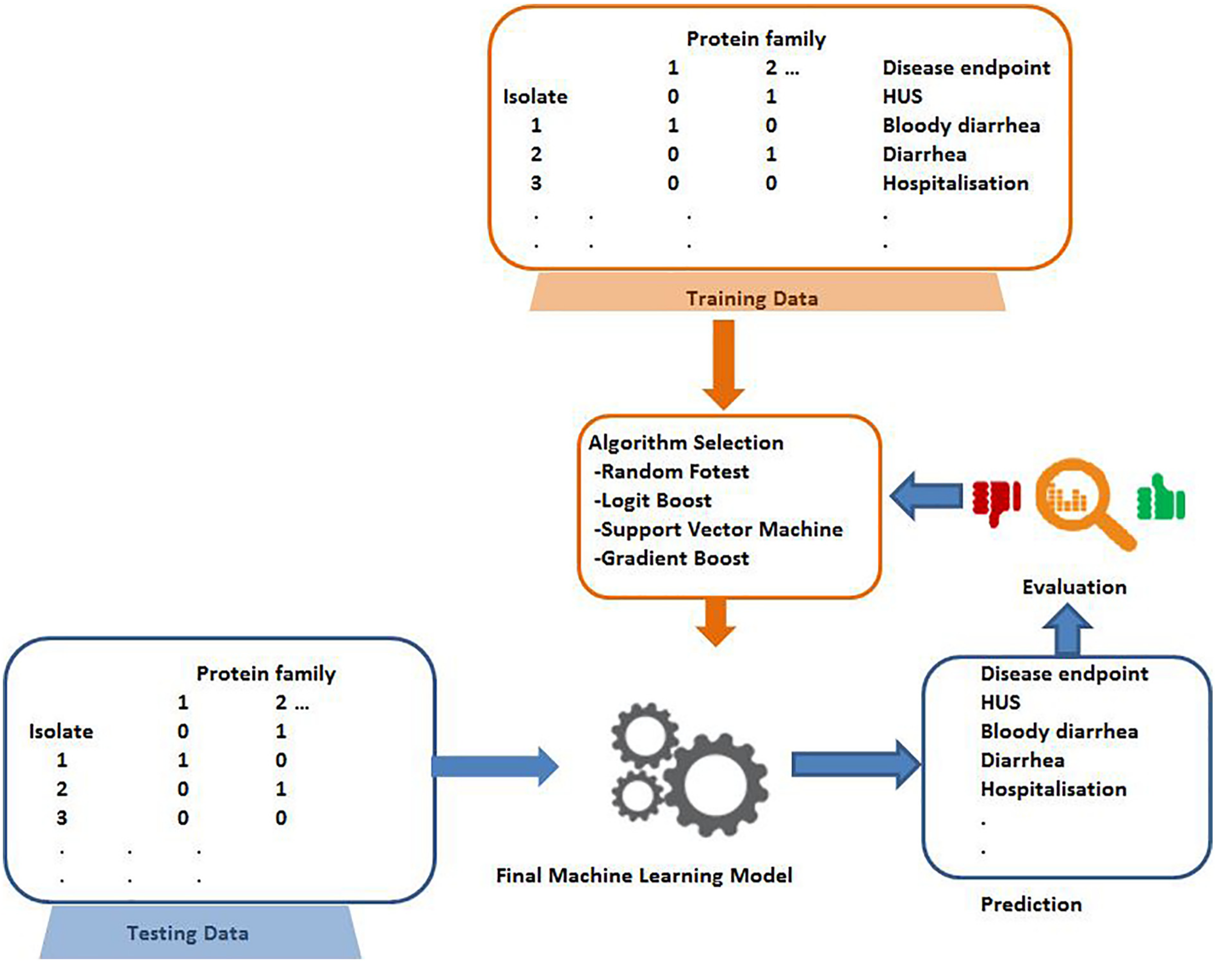

**Fig. 1.** Machine learning workflow for predicting STEC infection outcome using whole genome sequencing data. A training set of protein families consisting of data from a proportion of the isolates used as a machine learning input together with the clinical outcome (Hemolytic Uremic Syndrome (HUS), diarrhea, bloody diarrhea and hospitalization) associated with each isolate. A test set was used to evaluate the performance of the models with cross validation performed by splitting the data into 10 training and test sets followed by a repeat of model training and evaluation.

classification methods, ensemble methods have yielded more accurate models in many fields of applications (Zhou, 2012). Ensemble methods aggregate multiple weighted models so as to yield a unit model which outperforms every of the constituent single models (Ren et al., 2016). This family of models include methods such as bootstrap aggregation (bagging), adaptive boosting (boosting) and random forest, decomposition methods, negative correlation learning methods, multi-objective optimization based ensemble methods, fuzzy ensemble methods, multiple kernel learning ensemble methods and deep learning based ensemble methods (Ren et al., 2016).

### *2.3.1. Models*

Machine learning models were evaluated from algorithms that have been used in genetics including random forest (RF) (Machado et al., 2015; Ogutu et al., 2011), support vector machine (SVM) (radial and linear kernels) (Kuhn, 2008; Ogutu et al., 2011), and logit boost (LB) (Kuhn, 2008).

**Random forest** machine learning exhibits a number of appealing properties of potential interest for predictive risk modelling using WGS data including: (i) RF performs well in situations where number of features far exceed that of samples, (ii) it is robust and benefits from predictors showing weak effects, high correlations and interactions, (iii) has shown high accuracy for both simple and complex classification as well as regression problems, (v) has modest fine-tuning requirements for parameters, such that default parameterization is also adequate in many instances, and (vi) no assumptions are made about the distribution associated with the predictor variables (Ogutu et al., 2011).

**Support vector machine** (SVM) represent a class of powerful, highly flexible modelling techniques which are robust to outliers (Kuhn, 2008). The method applies kernel functions of inner products of predictors by arraying predictors in the observation space using a set of inner products (Hastie et al., 2009).

**Logit boost** (LB) and stochastic gradient boosting methods are part of a boosting family of algorithms which appeared in the early 1990s (Freund, 1995; Freund and Schapire, 1999; Schapire, 1990). In boosting, a number of weak classifiers (a weak classifier in one that predicts marginally better than random) are coalesced (or boosted) resulting in an ensemble classifier with a superior generalized classification accuracy (Kuhn and Johnson, 2013). Algorithms evolved beginning with AdaBoost and later on to Friedman's **stochastic gradient boosting** (GB), which has received extensive acceptance as the boosting algorithm of choice for machine learning applications. Like random forests, GB models process interactions effectively, are able to select variables automatically, are robust to outliers, missing data, many correlated as well as less important variables, and variable importance can be similarly generated.

All analyses were conducted in *R Version 3.2.3* and the dataset as well as the R code are presented in Supplemental material.

### *2.3.2. Data exploration*

The data were initially explored for zero-variance predictors as proposed by Kuhn and Johnson (2013). Such predictors have unique values at low frequencies and may further yield zero-variance predictors during subsequent splitting of the data into cross-validation/bootstrap subsamples. Zero-variance predictors also lead to model fit instabilities (Kuhn and Johnson, 2013).

### *2.3.3. Subsampling for class imbalances*

Exploration of clinical outcome classes showed considerable class imbalances (Supplemental Fig. 1). Such class imbalances may lead to models with poor overall class specific performance (Velez et al., 2007), because the model training process tends to be biased towards important patterns in the predictors associated with the larger classes. Class balance may not be achievable with WGS data as most projects depositing data do not set measures to a priori consider class balance of clinical outcomes since sampling is driven by the epidemiological situation. Post hoc sampling approaches have been proposed to mitigate the effects of the imbalance on the trained model (Kuhn and Johnson, 2013). The lowest class frequency had a considerably low number of samples and up-sampling was therefore chosen using an approach available in the R environment (Supplementary material and Supplemental Figs. 1 and 2). Logit boost learning method was used for comparison of resampling effectiveness with 10 cross-validations.

### *2.3.4. Data splitting*

Data were divided into training (70%) and testing sets (30%) (Fig. 1). Resampling was also performed by 10 times cross-validation using multiple alternate versions of the train and test datasets. The possibility of overfitting in machine learning models is checked by cross-validation and the model's valid accuracy where the accuracy scores decrease if there is overfitting and only “valid accuracy” is used. “Out of bag error” estimates (OOB) help safeguard against overfitting for tree based methods. OOB is the mean prediction error on each training sample $x_i$, using only the data that did not have $x_i$ in their bootstrap sample (Kuhn and Johnson, 2013). For each random sampling during a model run, a sample is setaside and is used to assess errors. Multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance. Cross-validation was performed by randomly partitioning model input samples into 10 sets of roughly equal size followed by estimation of accuracy based on held-out samples. This held-out sample was returned to the training set each time and the procedure was repeated with the second subset held out and so forth (Kuhn and Johnson, 2013).

### *2.3.5. Model selection*

Machine learning models RF, SVM (radial and linear kernels) and LB were evaluated. Different techniques possess potentially useful characteristics depending on the type of data and the methods were evaluated for predictive performances.

Models were built with 10 times cross validation by random data splitting, training of the models, making predictions and recording of accuracies after each run using the caret packages for the R statistical environment (Kuhn et al., 2012; Liaw and Wiener, 2002) (Supplementary material). Cross-validation and parallel processing were enabled by the inclusion of a train control parameter.

Analysis of variance, at significance alpha value of 0.05 was used to analyze the differences in mean accuracy between the models.

### *2.3.6. Model evaluation*

A confusion matrix was plotted as an initial model accuracy check by cross-tabulating observed and predicted classes. The confusion matrix describes the performance of a classification model on a set of test data for which the true values are known. Accuracy scores are also calculated from the confusion matrix as:

$$(\text{True Positive} + \text{True Negative})/\text{Total}$$

Accuracy depicts the agreement between the observed and predicted classes. Posterior distribution of the balanced accuracy (balanced accuracy) rather than the use of average accuracy was used to calculate the accuracy over the 10 fold cross-validations. The average accuracy approach does not yield meaningful confidence intervals and also leads to an optimistic estimate when a biased classifier is tested on an imbalanced dataset (Brodersen et al., 2010). Cohen's Kappa was used for further inference from the confusion matrix diagnosis for the accuracy of class distributions. Kappa statistic values range from −1 to 1 such that zero values imply no agreement between the observed and predicted classes, whereas values of 1 suggest perfect agreement. Two of the proposed cut-off values were used to make conclusions about the model accuracies. These include Kappa statistic values of “0–0.20 = slight”, “0.21–0.40 = fair”, “0.41–0.60 = moderate”,

**Table 1**
Model performance for imbalanced data and after subsampling for STEC outcome class imbalances using up-sampling.

| Class accuracy | Imbalanced[a] | Upsampled[b] |
|---|---|---|
| D | 0.55 | 0.69 |
| D_BD | 0.38 | 0.67 |
| D_BD_H | 0.51 | 0.92 |
| D_BD_HUS_H | 0.50 | 0.99 |
| D_H | 0.50 | 1.00 |
| Average accuracy | 0.28 (0.14–0.47) | 0.78 (0.64–0.89) |
| Kappa | −0.05 | 0.72 |

D - diarrhea, BD - bloody diarrhea, H - hospitalization, HUS - hemolytic uremic syndrome, CI - confidence interval.

[a] Model from original data with unequal number of isolates per class.

[b] Model after mitigation for this class imbalance by up-sampling cases from the minority classes with replacement until each class has approximately the same number.

“0.61–0.80 = substantial”, and “0.81–1 = almost perfect” as proposed by Landis and Koch (1977). Another option suggested by Fleiss et al. (2003) involves a description of models with Kappa values $>$ 0.75 as excellent, 0.40–0.75 fair to good, and $<$ 0.40 as poor. Prediction accuracy of clinical outcome categories was assessed based on sensitivity, specificity, positive predictive value and negative predictive value. The positive predictive value is the probability that subjects with a positive screening test truly have the disease while the negative predictive value provides the probability that subjects with a negative screening test are truly disease negative (Altman and Bland, 1994).

### 2.3.7. *Variable importance*

Gene families important in predicting clinical outcomes were selected. Variable or feature selection identifies unnecessary, irrelevant and redundant features from data that either make negligible contributions to the accuracy of a predictive model or may even decrease the accuracy of the model. Use of fewer predictors is advantageous because it reduces model complexity. The aim of important gene selection was to allow selection of predictors yielding improved model performance and efficiency. These predictors may be genes associated with certain clinical outcomes and epidemiological inference concerning single isolates provides a better understanding of the underlying virulence and stress response. Model-based rather than outside variable importance measures allow better performing models to be used for variable selection while incorporating the correlation structure between the predictors in variable selection. However, over-pruning of the input data set may result in seclusion of some relevant features. We, therefore, compared the model based important feature selection from logit boost with that of two other approaches. These approaches included Boruta algorithm (Kursa, 2014) which find all features which are either strongly or weakly relevant to the decision variable. Feature selection was also performed using a recursive feature elimination method involving backwards selection followed by cross-validation. Proteins represented by important genes were predicted by blasting their amino acid sequences in BLASTP (NCBI) (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) followed by comparison of predictions with those from the database Uniprot (http://www.uniprot.org).

## 2.4. *Data availability*

Identification codes and accessory protein amino acid sequences of the protein families used as input are provided in Supplemental table and Supplemental data.

## 3. Results

### 3.1. *Clinical outcomes and strain characteristics*

Out of the initial 5737 protein clusters, 4255 were near zero variance predictors and therefore a final 1482 were selected for further modelling. The associations between clinical outcome and MLST type, lineage specific polymorphism assay (LSPA-6), stx-subtype and sub-lineage type were assessed based on Pearson's chi-squared test ($\chi^2$). No significant association was found between clinical outcome and MLST type ($p$-value = 0.68), LSPA-6 ($p$-value = 0.18), stx-subtype ($p$-value = 0.52) and sub-lineage type ($p$-value = 0.29).

The association between clinical outcomes and the covariates age group and travel was also assessed. Age groups were divided into children (< 10 years), youth (10–17 years), adults (18–59 years) (Keithlin et al., 2014) and elderly (≥60 years). Clinical outcome was significantly associated with age group ($\chi^2$ = 25, 12 df, $p$-value = 0.015). Due to the low frequency of travel versus clinical outcome tabulated values, country specific travel cases were not considered and classes were collapsed into travel and UK domestic cases. There was a borderline significant association between travel and clinical outcome ($\chi^2$ = 9.59, 4 df, $p$-value = 0.048).

### 3.2. *Predictive modelling*

The initial model with class imbalances (Supplemental Fig. 1) performed poorly at an accuracy of 0.28 (CI: 0.14, 0.47) and dismal Kappa value of −0.05. Mitigation for this class imbalance was performed by up-sampling cases from the minority classes with replacement until each class had approximately the same number.

Table 1 shows results from the average accuracy for the 10 cross-validations comparing both the original and up-sampled data. The results indicate that up-sampling significantly improved the model performance to an accuracy of 0.78 (CI: 0.64, 0.89). A Kappa value of 0.72 indicated that this model performed substantially well according to criteria by Landis and Koch (1977) and excellent according to the criteria proposed by Fleiss et al. (2003).

#### 3.2.1. *Model selection*

We compared the performances of the machine learning methods random forest (RF), support vector machine (SVM) (radial and linear kernels) and logit boost (LB). Model evaluation was based on average accuracy from 10 cross-validations for the candidate models (Fig. 2).

Logit boost was the best performing model followed by SVM-linear and RF respectively and these differences were statistically significant (F-statistic: 3.7 on 4 DF, $p$-value: 0.01). Tukey multiple comparisons of the mean accuracies, however, indicated that only SVM-linear versus LB and SVM-linear versus SVM-radial were significantly different from each other ($p < 0.05$). LB was chosen for further inference and the agreement accuracy between LB and the other models were 0.81, 0.81, 0.83 and 0.81 for Random Forest, SVM-linear kernel and SVM-radial kernel respectively.

#### 3.2.2. *Final logit boost model*

The final LB model was trained using 70% of the data and tested using the rest of the data (30%) and 10-fold cross-validation was applied to produce performance estimates. The accuracy of the model was 0.75 (95% confidence interval: 0.60–0.86) which was achieved after 11 iterations.

The Kappa statistic was 0.69, which is substantial according to criteria by Landis and Koch (1977) and fair to good according to Fleiss et al. (2003). Sensitivity, specificity, positive predictive, and negative predictive values were all ≥0.8 except for the outcomes diarrhea and bloody diarrhea, where there were lower values in some performance measures (Table 2).

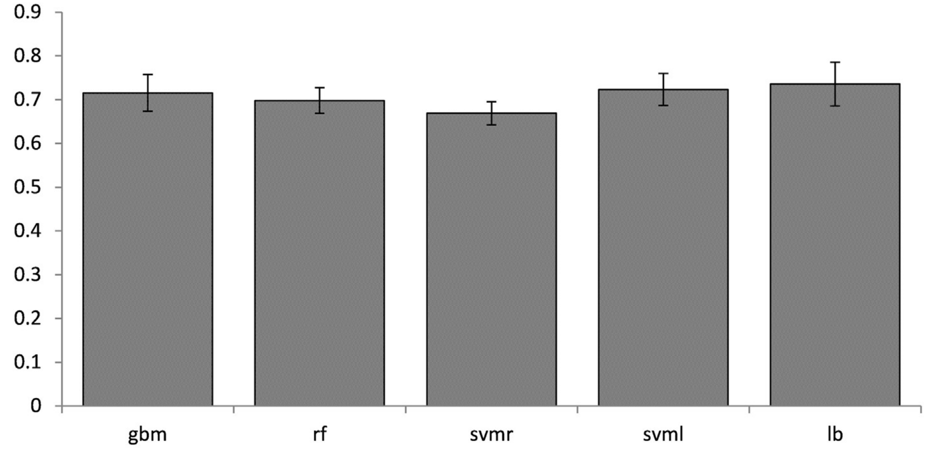

**Fig. 2.** Predictive performances from 10 cross-validations of random forest (rf), support vector machine (radial (svmr), Gradient Boosting (gbm) and linear kernels (svml), and logit boost (lb) models. Error bars represent standard deviations.

**Table 2**
Model performance for shigatoxigenic *Escherichia coli* clinical outcome predictions from the logit boost model.

| | Class | | | | |
|---|---|---|---|---|---|
| | D | D_BD | D_BD_H | D_BD_HUS_H | D_H |
| Sensitivity | 0.50 | 0.38 | 0.89 | 1.00 | 1 |
| Specificity | 0.87 | 0.95 | 0.95 | 0.97 | 1 |
| PPV[a] | 0.25 | 0.71 | 0.80 | 0.93 | 1 |
| NPV[b] | 0.95 | 0.81 | 0.98 | 1.00 | 1 |
| Balanced accuracy | 0.68 | 0.67 | 0.92 | 0.99 | 1 |

D - diarrhea, BD - bloody diarrhea, H - hospitalization, HUS - hemolytic uremic syndrome.

[a] Positive predictive value.

[b] Negative predictive value.

### 3.2.3. Important predictor proteins

Most important predictor protein families included A0747, A0253, A0259, A5715, A2240, A0434, A0702, A0710, A0712, A0882, A0899, A0925, A0942, A3466, A3764, A4831, A4856, A0508, A0898, A0932 and A0960 (Fig. 3). The amino acid sequences corresponding to these identification codes are provided in Supplemental table. Patterns of the probabilities of the respective proteins in predicting a certain class uniquely could be distinguished (Fig. 3) and we discuss here genes uniquely predicting specific clinical outcome categories at probabilities > 0.8. The proteins A0253, A0259 and A0747 predicted the outcomes diarrhea with hospitalization or bloody diarrhea without hospitalization. The protein A5715 predicted the outcome bloody diarrhea with hospitalization and the sequela HUS with hospitalization. The rest of the top predictor proteins namely A0434, A0508, A2240, A0702, A0710, A0712, A0882, A0899, A0925, A0932, A0942, A0960, A3466, A3764, A4831 and A4856 predicted bloody diarrhea accompanied by hospitalization (Fig. 3).

Due to the significance of the association between travel and clinical outcome, we assessed the prediction probabilities of either travel or domestic cases by the top predictor proteins (Section 3.1). The proteins A0259, A0253 and A0747 which were also among the top four most important predictors were associated at high probabilities with travel cases (Supplemental Fig. 2).

Table 3 shows the most important predictor proteins and their predicted identity or biological function, where this is known. Among these 21 proteins, the roles of five are not defined and furthermore the role that some of the other proteins play in virulence and therefore enhanced risk was not clear.

## 4. Discussion

Microbial risk assessment has been widely incorporated as a scientific basis for deriving measures for public health protection. The risk can be prospectively estimated and many countries have adopted risk based microbial criteria and legislation. It is common when conducting hazard characterization during risk assessment to consider the species as a homogeneous virulence unit when deriving parameters that define dose-response. For instance, in the case of STEC, varying outcomes and severity such as diarrhea, severe diarrhea, hemolytic uremic syndrome (HUS) and other chronic sequelae have not been well distinguished in MRA (Franz et al., 2014; Spinale et al., 2013). The heterogeneity of risk posed by STEC is well known from the diverse population structure and the associated emergence of non-O157 STEC which have been associated with an increasing number of outbreaks. This complicates the predictive accuracy of risk assessment efforts. WGS data provides an opportunity to harness this variation in pathogenicity between microbial strains from differing sources (Pielaat et al., 2013). One major hurdle in translating microbial genotypic data to phenotypic clinical outcomes lies in the high dimensional nature of the data and a complex interaction between genetic factors that define disease outcomes. Approaches such as single-variant association testing and GWAS studies have been utilized in human disease to decipher the genetic variation leading to particular traits and human disease. However these methods have left a vast portion of the heritability unexplored and the clinical utility of proposed individual and combined effects is still diminutive (Maher, 2008; Okser et al., 2013). Machine learning methods provide an opportunity in hazard characterization and risk prediction for the interpretation of such large and complex data sets. This is because machine learning techniques ‘learn’ to recognize important patterns in the data (Libbrecht and Noble, 2015). This study proposes machine learning as a hazard characterization approach in support of WGS based microbial risk assessment by the accurate prediction of the STEC clinical endpoints including diarrhea, bloody diarrhea and HUS or their combinations with isolate sequence data as input.

A diverse sequence dataset from a diverse panel of STEC in terms of MLST types, lineage specific polymorphisms (LSPA-6), stx-subtypes and sub-lineage types from both sporadic cases and outbreaks was used to train and evaluate machine learning models for their predictive potential on clinical outcome. This resulted in a fairly accurate (Accuracy of 0.75 with 95% CI: 0.60, 0.86) final LB model. Incorrect classifications are often attributable to a diverse genetic population structure where certain microbial strains associated with diverse disease phenotypes are not accounted for (Okser et al., 2013; Tian et al., 2008). There is, therefore, opportunity for classification improvement as sequencing efforts yield more strains with further diversity. However, the

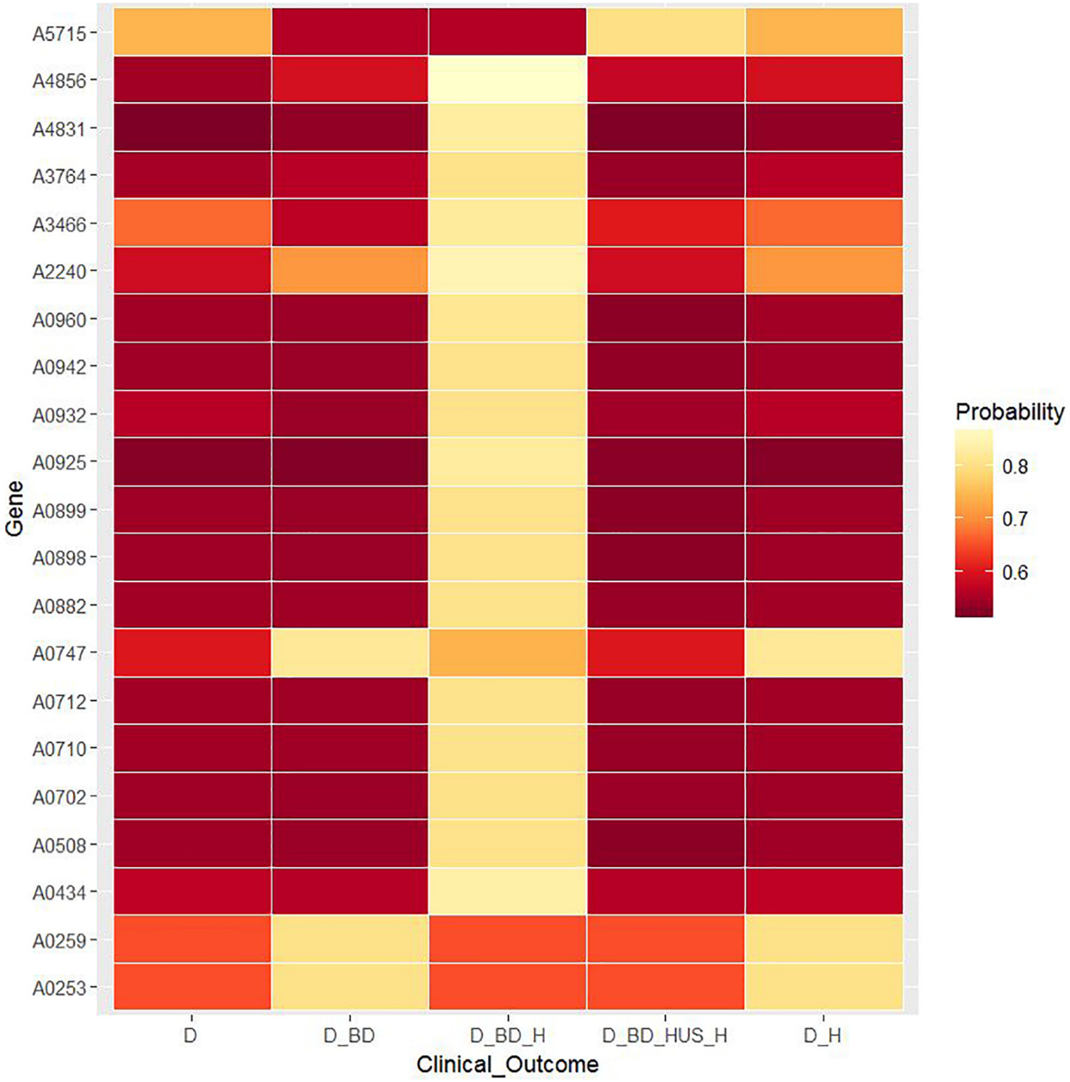

**Fig. 3.** Twenty one most important predictor proteins for Shigatoxigenic Escherichia coli clinical outcome presented by their relative importance and class probabilities.

D- diarrhea, BD- bloody diarrhea, H -hospitalization, HUS- hemolytic uremic syndrome.

**Table 3**

Biological information regarding the top 21 predictor proteins for clinical outcomes.

| ID[a] | Predicted protein | Predicted protein organism | Predicted protein accession number |
|---|---|---|---|
| A0747 | Tail fiber protein | *Escherichia coli* O157:H7 str. K1793 | EZB52764.1 |
| A0253 | mRNA endoribonuclease LsoA | *Escherichia coli* | WP_089564644.1 |
| A0259 | Antitoxin LsoB | *Escherichia coli* | WP_032345734.1 |
| A5715 | NAD-dependent malic enzyme | *Escherichia coli* DORA_A_5_14_21 | ETJ28701.1 |
| A2240 | Hypothetical protein ECH7EC4113_1978 | *Escherichia coli* O157:H7 str. EC4113 | EDU53199.1 |
| A0434 | Hypothetical protein ECTX1999_1477 | *Escherichia coli* TX1999 | EGX24327.1 |
| A0702 | Cobalamin biosynthesis protein CbiX | *Escherichia coli* | WP_097178501.1 |
| A0710 | traF protein | *Escherichia coli* | WP_095526288.1 |
| A0712 | Toxin co-regulated pilus biosynthesis Q family protein | *Escherichia coli* | WP_021503160.1 |
| A0882 | Hypothetical protein | *Escherichia coli* | WP_063106769.1 |
| A0899 | Conjugal transfer prepropilin | *Escherichia coli* O157:H7 str. EC4401 | EDU72803.1 |
| A0925 | Type II toxin-antitoxin system HicA family toxin | *Escherichia coli* | WP_074180779.1 |
| A3466 | Hypothetical protein ECDEC4C_1418 | *Escherichia coli* DEC4C | EHV11465.1 |
| A3764 | Type III secretion system LEE transcriptional regulator GrlA | *Escherichia coli* | WP_000444180.1 |
| A4831 | Acetyltransferase, partial | *Escherichia coli* O157:H7 str. K2191 | EZB98504.1 |
| A4856 | Putative transposase for insertion sequence element, partial | *Escherichia coli* 2-156-04_S4_C2 | KDX30686.1 |
| A0508 | Type IV secretory system Conjugative DNA transfer family protein | *Escherichia coli* FRIK1999 | EKH26420.1 |
| A0898 | CopG family transcriptional regulator | *Escherichia coli* | WP_085445835.1 |
| A0932 | Conjugal transfer protein TraD (plasmid) | *Escherichia coli* O157:H7 str. EC4115 | ACI39845.1 |
| A0960 | Hypothetical protein EC970246_A0059 | *Escherichia coli* 97.0246 | EIG93754.1 |

[a] The complete list of proteins and their sequences is described in Supplemental Table 1.

bottleneck remains the scarcity of data on sequenced strains accompanied by a well-defined set of clinical outcomes.

Logit boost was selected as the best performing model both in accuracy and agreement with other comparatively well performing models. LB models render themselves particularly attractive choice of learners due to a number of inherent theoretical and algorithmic features (Ferreira and Figueiredo, 2012). LB is part of the boosting class of machine learning methods which rely on combination of weak classifiers to yield classifiers that perform better than the single classifiers. In LB, adaptive Newton steps are used to fit an additive logistic model where the logistic loss is minimized instead of minimization of the exponential loss function (Ferreira and Figueiredo, 2012; Friedman et al., 2000). Davis et al. (2016) recently used AdaBoost, a boosting algorithm, for accurate prediction of carbapenem resistance in *Acinetobacter baumannii*, methicillin resistance in *Staphylococcus aureus*, and beta-lactam and co-trimoxazole resistance in *Streptococcus pneumoniae*.

Although the presence of certain biomarkers constitutes only one of a large repertoire of possible risk predictors, machine learning models support the selection of subsets of features which may explain the differences in disease risk outcomes. Such features constitute a subset with best predictive potential and may increase our understanding of the biological background of virulence outcomes (Glaab et al., 2012; Libbrecht and Noble, 2015; Urbanowicz et al., 2012).

The four proteins A0747, A0253, A0259 and A3093 predicted the occurrence of diarrhea with hospitalization or bloody diarrhea alone or bloody diarrhea accompanied by hospitalization. The predicted protein for A0747 is a putative tail fiber protein encoded by gene *SS52_0295* which is associated with the tail fiber protein in *Escherichia coli* O157:H7 str. K1793. Enterobacteria phage proteins are a structural component of the short non-contractile tail which may attach to host lipopolysaccharides (LPS) therefore facilitating initial attachment to the host cell. Predicted proteins for A0253 and A0259 were LsoA (encoded by gene *LsoA*) and LsoB (encoded by gene *LsoB*) respectively. These proteins are part of the type II toxin-antitoxin (TA) system whose co-expression is essential for cell survival. These genes co-occurred in 49 (47%) of the isolates and did not occur alone in any of the isolates. Overexpression of LsoA without LsoB leads to retarded cell growth and mRNA degradation. The TA system contributes to the selective persistence of plasmids or genomic islands, including super-integrons, as a result of the post-segregational death of a cell that loses these genes thereby exposing the cell to destruction by the stable toxin (Van Melderen and De Bast, 2009). Toxin-antitoxin systems are potential stress adaptation markers which may mediate diversity in environmental persistence of microbial strains.

The protein A5715 predicted the outcome HUS in addition to bloody diarrhea and hospitalization. The predicted protein was NAD-dependent malic enzyme encoded by gene *maeA*. It is important to further elucidate the mechanism through which this protein contributes to HUS especially because the results indicated this protein predicts HUS. The STEC infection sequela HUS is known to be common for the very young and old (Spinale et al., 2013). In this dataset, HUS was however reported for all age groups. A test of association between age and protein A5715 showed no significant association between STEC containing this protein and the age of patients with HUS ($\chi2 = 4.44$, 3 df, $p\text{-value} = 0.22$, $\text{power} = 0.73$ for a medium effect size of 0.3 (Cohen, 1988)) which points to a general association with HUS for all age groups. However, this needs to be evaluated in future epidemiological studies with larger sample sizes.

The rest of the proteins A0434, A0508, A2240, A0702, A0710, A0712, A0882, A0898, A0899, A0925, A0932, A0942, A0960, A3466, A3764, A4831 and A4856 predicted bloody diarrhea accompanied by hospitalization. A0702 predicts cobalamin biosynthesis protein CbiX and occurred in 49 (47%) of the strains. The role of cobalamin in host infection may be related to the pivotal and well known role of iron in infection. Both host and pathogen undergo a series of shared complex changes in iron and vitamin $B_{12}$ during infection. In *E. coli*, a similar transport system for iron and vitamin $B_{12}$ and similar source of binding proteins for both are indicative of the possible role of cobalamin in *E. coli* pathogenicity (Neale, 1990). However, previous studies using infant-rat and chicken embryo models indicated no difference in virulence between a cobalamin receptor deficient mutant *E. coli* strain and a wild type (Sampson and Gotschlich, 1992). The higher bloody diarrhea and hospitalization shown by strains having cobalamin biosynthesis protein in our study suggest a possible important role in human STEC infections of *E. coli* expressing this cobalamin biosynthesis protein (Sampson and Gotschlich, 1992). Studies in *Salmonella enterica* serovar Gallinarum indicated that a mutant defective in cobalamin biosynthesis was half as virulent as the wild type in chickens (de Paiva et al., 2009). The protein A0710 was predicted as traF protein which is encoded by gene *BX52_25085*. TraF is encoded by *E. coli* F plasmid and plays a role in conjugative plasmid transfer and formation of sex pili (Audette et al., 2004). A0712 corresponded to the toxin co-regulated pilus biosynthesis Q family protein of *E. coli*. This toxin connected to the pilus biosynthesis family is also reported as an indicator for pathogenicity in *V. cholerae* and may be part of the toxin repertoire leading to diarrhea gained by *V. cholerae* from *E. coli* (Georgiades and Raoult, 2011). A0899 was predicted as the conjugal transfer prepropilin from *Escherichia coli* O157:H7 str. EC4401 encoded by the gene *CEP72_29710*. A0925 was predicted as type II toxin-antitoxin system HicA family toxin encoded by *Escherichia coli* gene *CEP72_29820*. Like LosA and LosB, this is one of the common toxin-antitoxin (TA) systems in bacteria composed of both a stable "toxin" component and an unstable "antitoxin" (Jurenaite et al., 2013). Whereas the toxins consist of proteins, the antitoxin is either RNA (TA types I and III) or a protein (TA types II, IV, and V) (Jurenaite et al., 2013). This antitoxin may, therefore, contribute to the survival of the *E. coli* cells through plasmid stabilization as well as guarding against toxin related growth retardation and cell death under stressful conditions. This may mediate environmental persistence of strains and successful passage after ingestion to the site of infection, which leads to variations in exposure estimates in MRA.

A3764 was predicted as type III secretion system LEE transcriptional regulator GrlA. Several regulatory elements control the gene function of the locus of enterocyte effacement which is important for virulence in pathogenic *Escherichia coli* especially with respect to bloody diarrhea. The down-regulation of intracellular levels of GrlR, a negative regulator of LEE gene expression is mediated through GrlA, a positive regulator of LEE expression (Iyoda et al., 2006). This role of the protein GrlA was apparent in the importance of this protein as a predictor of bloody diarrhea and hospitalization in our study.

A4831 was predicted as acetyltransferase, from *Escherichia coli* O157:H7 str. K2191. The *E. coli* N$\epsilon$-acetyltransferase (PatZ) is the only enzyme reported to catalyze the post-translational acetylation of proteins where *E. coli* PatZ is uniquely acetylated in vivo with unknown consequences (De Diego Puente et al., 2015). The most well-known PatZ substrate is acetyl-CoA synthetase which is regulated by acetylation in bacteria such as *S. enterica*, *E. coli*, *Bacillus subtilis*, *Rhodopseudomonas palustris* and Mycobacterium tuberculosis most of which are important pathogens (De Diego Puente et al., 2015). The role of this protein in bloody diarrhea accompanied by hospitalization in STEC may be of further interest.

A4856 was predicted as the putative transposase for insertion sequence element in *Escherichia coli* 2-156-04_S4_C2. Insertion sequences (ISs) are ubiquitous and abundant mobile genetic elements in prokaryotic genomes normally responsible for the encoding of a single protein referred to as the transposase which acts as a catalyst for their transposition (Díaz-Maldonado et al., 2015). The role of IS elements in the evolution of pathogenicity can be attributed to their facilitation of the rearrangement or deletion of sections within the pathogenic islands which lead to the evolution of new variant strains and promotes strain adaptation (Hallstrom and McCormick, 2014).

A0508 was predicted as a protein from the type IV secretory system conjugative DNA transfer family. The type IV secretion system (T4SS) is a part of a group of secretion systems used by microorganisms to transport macromolecules such as proteins and DNA across the cell envelope (Wallden et al., 2010). Pathogenic Gram-negative bacteria such as *E. coli* use some T4SSs to translocate a myriad of virulence factors into the host cell as well as to transfer genes that enhance environmental adaptation and antimicrobial resistance (Wallden et al., 2010). A0898 was predicted as CopG family transcriptional regulator which regulates the plasmid copy number by binding to the RepAB promoter. Finally, the protein A0932 was predicted as a conjugal transfer protein TraD (plasmid) of *Escherichia coli* O157:H7 str. EC4115. It has been proposed that the F sex factor TraD protein supports DNA transfer during conjugation since although TraD mutants carry out conjugation successfully, they are unable to transfer DNA (Panicker and Minkley, 1992). Conjugative transfer supports adaptive evolution by facilitating DNA transfer which may alter pathogenicity over a rather short evolutionary distance in many broad-host-range plasmids in bacterial pathogens (Seubert et al., 2003). These genetic elements demonstrate the potential or WGS based MRA in the study of genetic elements associated with increased virulence and other important phenotypes that could be transferred between strains. This contributes to the dynamic nature of the microbial pathogen response to exposure and infection leading to changes in the definition of microbial hazards in MRA which can only be adequately captured by incorporating WGS as input for classical MRA.WGS also presents an opportunity for predicting the source of the more important pathogen variants. The proteins A0259, A0253 and A0747 were associated at high probabilities with travel cases (Supplemental Fig. 2). This confirms the possible role of the co-occurrence of the type II toxin-antitoxin (TA) system components LsoA and LsoB on stress adaptation of STEC in diverse environments. Such *E. coli* maintain adaptive plasmids or genomic islands, including super-integrons (Van Melderen and De Bast, 2009). Further approaches such as pathway analysis are recommended to infer if related loci in the same biological pathway may jointly predict interesting traits such as preservation stress survival, growth and/or virulence which are of potential importance when performing microbial assessment risk assessment along the food production chain (Okser et al., 2013). Such predictive models linking genotypes to stress adaptation and virulence will contribute to reduced requirements for laboratory and food matrix model validations (Okser et al., 2013).

We also propose the role of the approach reported here in improving hazard characterization in MRA. Past MRA has relied on historical strains from selected cases to define and characterize hazards. This assumes that the pathogen is a unit and neglects within-species heterogeneity in microbial virulence. In the context of classical risk assessment, the ML approach will support hazard identification by the application of next generation sequencing data as well as epidemiological data to derive higher resolution risk assessments. In MRA, dose–response modelling allows for the estimation of the probability of illness which depends on the concentration of ingested pathogenic microorganisms. Infection and subsequent illness occurs when a proportion of the ingested microorganisms survives human host barriers. The infection process consists of a number of steps such as survival and passage through the intestine, latching of *E. coli* on to the surface of an intestinal cell, injection of receptor proteins into the intestinal cell and formation of pedestal for bacterium by the intestinal cell leading to infection. Most current hazard characterization is conducted under the assumption that each ingested microorganism is a taxonomic unit that has the same probability to provoke illness. The number of microorganisms surviving different barriers in the host is assumed to follow a binomial distribution. However genes are capable of transfer between bacterial species thus adding heterogeneity in virulence within the taxonomic unit. The use of WGS approach defined in this study will involve taking the pathogen as a genetic unit for refined dose-response assessment. We propose that the pathogen is a genetic unit or strain *i*, which has a probability *pi* expressible as $pi = f(p1i, p2i\ldots pni)$ where each *pxi* is the probability of a strain *i* completing each of the *n* infection steps *x* for $x = 1, 2, \ldots, n$. This concept may be implemented by calculating *pi* for every *i* in the taxonomic unit population. This will indeed contribute to the redefinition of dose-response relationships for initial infection from the relative proportions of each strain in a WGS dataset. Classical hazard characterization efforts can therefore be complemented with inputting whole genome sequence data to make more refined clinical endpoint estimations. Pielaat et al. (2013) proposed an 'organization principle' where risk assessment based on sequencing data is grounded upon prioritizing highly pathogenic strains. This study provides such a link between genotypic and phenotypic properties as an approach to identify *high risk* or priority isolates from the full spectrum of strains. ML approach is well suited for both prediction as well as interpretation based on such large, complex and highly dimensional data sets, where machine learning techniques 'learn' to recognize important patterns in the data (Libbrecht and Noble, 2015). The logit boost model showed a high outcome prediction accuracy (0.75, 95% CI: 0.60, 0.86; Kappa = 0.72) and is therefore of potential as a more specific hazard characterization tool enabling the prediction of the STEC clinical endpoints including diarrhea, bloody diarrhea and HUS or their combinations given sequence data from an isolate of an unknown clinical outcome. We hypothesize that this approach will also support the setting of product specific microbial criteria which: (i) avoids the blanket removal of foods based on the findings of pathogens whose perceived threat is generalized as well as a redefinition of the potentially dynamic qualified presumption of safety (QPS) status of existing species (Pielaat et al., 2013), and (ii) provides more specific hazard identification and characterization which may provide important real-time resolution and prevention of outbreaks caused by foodborne bacteria. Another opportunity is to use ML techniques in the improvement of exposure assessment in MRA. Exposure assessment involves the study growth, survival or death of microorganisms as a function of the growth environment such as food and other environmental conditions from farm to fork. The results from this study indicate variations in STEC genetic pattern that may influence environmental stress resistance. Food safety concern is increased when such adaptation of microorganisms to changes in environments increases resistance to environmental, processing and host stress agents (Abee et al., 2004). With collection of phenotypic data such as adaptation to various environmental stresses such as salt, acid, desiccation and temperature, models can be trained to produce strain specific categorization into different stress response categories based on WGS data. With increase in such data, predictive models based on WGS will form a predictive platform for survival and eventual exposure. However, incorporation of the proposed approach in quantitative microbial risk assessment will require further inference from a larger collection of strains. This includes data supporting generalizability of the outcome classes, prevalence of the isolates associated with different outcome classes, and sample size considerations as to what constitutes an optimal number of strains representative of the molecular variation in pathogen population. Availability of WGS and clinical outcome data from a collection of specific pathogen and food production chains including prevalence, concentrations and food production chain properties is needed to conduct quantitative risk assessment capturing variability and/or uncertainty. A first important step will be the collection of a database consisting of bacterial genomes with stress response and clinical outcome metadata in order to advance exposure assessment and hazard characterization as well as the identification of genomic regions encoding different outcomes. The dataset used in the current study as well as R code are presented in the Supplemental material for use in prediction of clinical outcomes in *E. coli* and for future model building as further data from *E. coli* and other food borne pathogens become available.

## 5. Conclusions

An approach to improve precise hazard identification and prediction of specific clinical outcomes in STEC which would lead to improved inference from MRA by using WGS data was proposed. Furthermore, important characteristics that distinguish high from less risk strains were outlined. These characteristics include those mediating initial attachment to the host cell, persistence of plasmids or genomic islands, conjugative plasmid transfer and formation of sex pili, regulation of LEE expression, post-translational acetylation of proteins, facilitation of the rearrangement or deletion of sections within the pathogenic islands and transport macromolecules across the cell envelope. Further investigations regarding the involvement of the genes encoding these proteins and those with undefined functionality are recommended.

We foresee the increasing utility of the machine learning methods in microbial risk assessment, prediction and source tracking as more WGS data accompanied by clinical outcomes becomes available. This will expedite the detection of new pathogenic threats, which is an important prerequisite in reducing reaction times prior to and during outbreaks. Like in many areas where machine learning methods have found application, online WGS risk assessment tools incorporating the approach demonstrated in this study may lead to improved food safety and reduction of unnecessary product withdrawals, where non/low-pathogenic strains are involved.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijfoodmicro.2018.11.016.

## Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 643476 (http://www.compare-europe.eu/).

## Author contributions

P.M.K.N. and T.H. contributed to the study design, data modelling and writing the manuscript; P.L. contributed with the downloading of the sequences and bioinformatics. All authors read and approved the manuscript.

## References

Abee, T., Van Schaik, W., Siezen, R.J., 2004. Impact of genomics on microbial food safety. Trends Biotechnol. 22, 653–660. https://doi.org/10.1016/j.tibtech.2004.10.007.

Altman, D.G., Bland, J.M., 1994. Diagnostic tests 2: predictive values. BMJ. https://doi.org/10.1136/bmj.309.6947.102.

Audette, G.F., Holland, S.J., Elton, T.C., Manchak, J., Hayakawa, K., Frost, L.S., Hazes, B., 2004. Crystallization and preliminary diffraction studies of TraF, a component of the *Escherichia coli* type IV secretory system. Acta Crystallogr. Sect. D: Biol. Crystallogr. 60, 2025–2027. https://doi.org/10.1107/S0907444904020724.

Binnewies, T.T., Hallin, P.F., Stærfeldt, H.H., Ussery, D.W., 2005. Genome update: proteome comparisons. Microbiology. https://doi.org/10.1099/mic.0.27760-0.

Breiman, L., 2001. Statistical modeling: the two cultures. Stat. Sci. 16, 199–215. https://doi.org/10.2307/2676681.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. In: Proceedings - International Conference on Pattern Recognition, pp. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.

Bruant, G., Maynard, C., Bekal, S., Gaucher, I., Masson, L., Brousseau, R., Harel, J., 2006. Development and validation of an oligonucleotide microarray for detection of multiple virulence and antimicrobial resistance genes in *Escherichia coli*. Appl. Environ. Microbiol. 72, 3780–3784. https://doi.org/10.1128/AEM.72.5.3780-3784.2006.

Brul, S., Bassett, J., Cook, P., Kathariou, S., McClure, P., Jasti, P.R., Betts, R., 2012. "Omics" technologies in quantitative microbial risk assessment. Trends Food Sci. Technol. 27, 12–24. https://doi.org/10.1016/j.tifs.2012.04.004.

Bugarel, M., Beutin, L., Fach, P., 2010. Low-density macroarray targeting non-locus of enterocyte effacement effectors (nle genes) and major virulence factors of Shiga toxin-producing *Escherichia coli* (STEC): a new approach for molecular risk assessment of STEC isolates. Appl. Environ. Microbiol. 76, 203–211. https://doi.org/10.1128/AEM.01921-09.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P., Van Eerdewegh, P., 2005. Identifying SNPs predictive of phenotype using random forests. Genet. Epidemiol. 28, 171–182. https://doi.org/10.1002/gepi.20041.

Buvens, G., De Gheldre, Y., Dediste, A., De Moreau, A.I., Mascart, G., Simon, A., Allemeersch, D., Scheutz, F., Lauwers, S., Piérard, D., 2012. Incidence and virulence determinants of verocytotoxin-producing *Escherichia coli* infections in the Brussels-Capital Region, Belgium, in 2008–2010. J. Clin. Microbiol. 50, 1336–1345. https://doi.org/10.1128/JCM.05317-11.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinforma. 10. https://doi.org/10.1186/1471-2105-10-421.

Carriço, J.A., Sabat, A.J., Friedrich, A.W., Ramirez, M., ESCMID Study Group for Epidemiological Markers (ESGEM), 2013. Bioinformatics in bacterial molecular epidemiology and public health: databases, tools and the next-generation sequencing revolution. Euro Surveill. https://doi.org/10.1038/jid.2013.1.Research.

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. Stat. Power Anal. Behav. Sci. https://doi.org/10.1234/12345678.

Coombes, B.K., Wickham, M.E., Mascarenhas, M., Gruenheid, S., Finlay, B.B., Karmali, M.A., 2008. Molecular analysis as an aid to assess the public health risk of non-O157 shiga toxin-producing *Escherichia coli* strains. Appl. Environ. Microbiol. 74, 2153–2160. https://doi.org/10.1128/AEM.02566-07.

Cooper, K.K., Mandrell, R.E., Louie, J.W., Korlach, J., Clark, T.A., Parker, C.T., Huynh, S., Chain, P.S., Ahmed, S., Carter, M., 2014. Comparative genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7. BMC Genomics 15, 17. https://doi.org/10.1186/1471-2164-15-17.

Davis, J.J., Boisvert, S., Brettin, T., Kenyon, R.W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A.R., Will, R., Xia, F., Stevens, R., 2016. Antimicrobial resistance prediction in PATRIC and RAST. Sci. Rep. 6, 27930. https://doi.org/10.1038/srep27930.

De Diego Puente, T., Gallego-Jara, J., Castaño-Cerezo, S., Sánchez, V.B., Espín, V.F., De La Torre, J.G., Rubio, A.M., Díaz, M.C., 2015. The protein acetyltransferase PatZ from *Escherichia coli* is regulated by autoacetylation-induced oligomerization. J. Biol. Chem. 290, 23077–23093. https://doi.org/10.1074/jbc.M115.649806.

de Paiva, J.B., Penha Filho, R.A.C., Arguello, Y.M.S., Berchieri Junior, A., Lemos, M.V.F., Barrow, P.A., 2009. A defective mutant of *Salmonella enterica* Serovar Gallinarum in cobalamin biosynthesis is avirulent in chickens. Braz. J. Microbiol. 40, 495–504. https://doi.org/10.1590/S1517-838220090003000012.

Díaz-Maldonado, H., Gómez, M.J., Moreno-Paz, M., Martín-Úriz, P.S., Amils, R., Parro, V., De Saro, F.J.L., 2015. Transposase interaction with the β sliding clamp: effects on insertion sequence proliferation and transposition rate. Sci. Rep. 5. https://doi.org/10.1038/srep13329.

Drouin, A., Giguère, S., Sagatovich, V., Déraspe, M., Laviolette, F., Marchand, M., Corbeil, J., 2014. Learning Interpretable Models of Phenotypes From Whole Genome Sequences With the Set Covering Machine.

Ferreira, A.J., Figueiredo, M.A.T., 2012. Boosting algorithms: a review of methods, theory, and applications. In: Ensemble Machine Learning: Methods and Applications, pp. 35–85. https://doi.org/10.1007/9781441993267_2.

Fleiss, J., Levin, B., Cho Paik, M., 2003. Statistical Methods for Rates and Proportions. John Wiley Sonshttps://doi.org/10.1198/tech.2004.s812. (1706, 800).

Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D.A., Bono, J., French, N., Osek, J., Lindstedt, B.A., Muniesa, M., Manning, S., LeJeune, J., Callaway, T., Beatson, S., Eppinger, M., Dallman, T., Forbes, K.J., Aarts, H., Pearl, D.L., Gannon, V.P.J., Laing, C.R., Strachan, N.J.C., 2014. 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. https://doi.org/10.1016/j.ijfoodmicro.2014.07.002.

Freund, Y., 1995. Boosting a weak learning algorithm by majority. Inf. Comput. 121, 256–285. https://doi.org/10.1006/inco.1995.1136.

Freund, Y., Schapire, R., 1999. Adaptive game playing using multiplicative weights. Games Econ. Behav. 29, 79–103. https://doi.org/10.1006/game.1999.0738.

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Stat. https://doi.org/10.1214/aos/1016218223.

Friis, C., Wassenaar, T.M., Javed, M.A., Snipen, L., Lagesen, K., Hallin, P.F., Newell, D.G., Toszeghy, M., Ridley, A., Manning, G., Ussery, D.W., 2010. Genomic characterization of *Campylobacter jejuni* strain M1. PLoS One 5. https://doi.org/10.1371/journal.pone.0012253.

Georgiades, K., Raoult, D., 2011. Comparative genomics evidence that only protein toxins are tagging bad bugs. Front. Cell. Infect. Microbiol. 1. https://doi.org/10.3389/fcimb.2011.00007.

Glaab, E., Bacardit, J., Garibaldi, J.M., Krasnogor, N., 2012. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PLoS One 7. https://doi.org/10.1371/journal.pone.0039932.

Gonzales, T.K., Kulow, M., Park, D.J., Kaspar, C.W., Anklam, K.S., Pertzborn, K.M., Kerrish, K.D., Ivanek, R., Döpfer, D., 2011. A high-throughput open-array qPCR gene panel to identify, virulotype, and subtype O157 and non-O157 enterohemorrhagic *Escherichia coli*. Mol. Cell. Probes 25, 222–230. https://doi.org/10.1016/j.mcp.2011.08.004.

Gould, L.H., Mody, R.K., Ong, K.L., Clogher, P., Cronquist, A.B., Garman, K.N., Lathrop, S., Medus, C., Spina, N.L., Webb, T.H., White, P.L., Wymore, K., Gierke, R.E., Mahon, B.E., Griffin, P.M., 2013. 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. https://doi.org/10.1089/fpd.2012.1401.

Griffith, O.L., Pepin, F., Enache, O.M., Heiser, L.M., Collisson, E.A., Spellman, P.T., Gray, J.W., 2013. A robust prognostic signature for hormone-positive node-negative breast cancer. Genome Med. 5, 92. https://doi.org/10.1186/gm496.

Hallstrom, K.N., McCormick, B.A., 2014. Pathogenicity islands: origins, structure, and roles in bacterial pathogenesis. In: Molecular Medical Microbiology, second edition. pp. 303–314. https://doi.org/10.1016/B978-0-12-397169-2.00016-0.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Springer, pp. 746. https://doi.org/10.1007/b94608. (2001, 18).

Havelaar, A.H., Brul, S., de Jong, A., de Jonge, R., Zwietering, M.H., ter Kuile, B.H., 2010. Future challenges to microbial food safety. Int. J. Food Microbiol. 139. https://doi.org/10.1016/j.ijfoodmicro.2009.10.015.

Holmes, A., Allison, L., Ward, M., Dallman, T.J., Clark, R., Fawkes, A., Murphy, L., Hanson, M., 2015. Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. J. Clin. Microbiol. 53, 3565–3573. https://doi.org/10.1128/JCM.01066-15.

Houle, D., Govindaraju, D.R., Omholt, S., 2010. Phenomics: the next challenge. Nat. Rev. Genet. 11, 855–866. https://doi.org/10.1038/nrg2897.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma. 11, 119. https://doi.org/10.1186/1471-2105-11-119.

Iyoda, S., Koizumi, N., Satou, H., Lu, Y., Saitoh, T., Ohnishi, M., Watanabe, H., 2006. The GrlR-GrlA regulatory system coordinately controls the expression of flagellar and LEE-encoded type III protein secretion systems in enterohemorrhagic *Escherichia coli*. J. Bacteriol. 188, 5682–5692. https://doi.org/10.1128/JB.00352-06.

Johnson, K.E., Thorpe, C.M., Sears, C.L., 2006. The emerging clinical importance of non-O157 Shiga toxin-producing *Escherichia coli*. Clin. Infect. Dis. 43, 1587–1595. https://doi.org/10.1086/509573.

Jurenaite, M., Markuckas, A., Suziedeliene, E., 2013. Identification and characterization of type II toxin-antitoxin systems in the opportunistic pathogen *Acinetobacter baumannii*. J. Bacteriol. 195, 3165–3172. https://doi.org/10.1128/JB.00237-13.

Karmali, M.A., Mascarenhas, M., Shen, S., Ziebell, K., Johnson, S., Reid-Smith, R., Isaac-Renton, J., Clark, C., Rahn, K., Kaper, J.B., 2003. Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. J. Clin. Microbiol. 41, 4930–4940. https://doi.org/10.1128/JCM.41.11.4930-4940.2003.

Keithlin, J., Sargeant, J., Thomas, M.K., Fazil, A., 2014. Chronic sequelae of *E. coli* O157: systematic review and meta-analysis of the proportion of *E. coli* O157 cases that develop chronic sequelae. Foodborne Pathog. Dis. 11, 79–95. https://doi.org/10.1089/fpd.2013.1572.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26. https://doi.org/10.1053/j.sodo.2009.03.002.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., 2012. Caret: classification and regression training. https://Cran.R-Project.Org/Package=Caret.

Kursa, M.B., 2014. Robustness of random forest-based gene selection methods. BMC Bioinforma. 15, 8. https://doi.org/10.1186/1471-2105-15-8.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159. https://doi.org/10.2307/2529310.

Leekitcharoenphon, P., Nielsen, E.M., Kaas, R.S., Lund, O., Aarestrup, F.M., 2014. Evaluation of whole genome sequencing for outbreak detection of *Salmonella enterica*. PLoS One 9. https://doi.org/10.1371/journal.pone.0087991.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22. https://doi.org/10.1177/154405910408300516.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332. https://doi.org/10.1038/nrg3920.

Machado, G., Mendoza, M.R., Corbellini, L.G., 2015. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. Vet. Res. 46, 1–15. https://doi.org/10.1186/s13567-015-0219-7.

Maher, B., 2008. Personal genomes: the case of the missing heritability. Nature 456, 18–21. https://doi.org/10.1038/456018a.

Neale, G., 1990. B12 binding proteins. Gut 31, 59–63.

Ogutu, J.O., Piepho, H.-P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 5 (Suppl. 3), S11. https://doi.org/10.1186/1753-6561-5-S3-S11.

Okser, S., Pahikkala, T., Aittokallio, T., 2013. Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. BioData Min. 6, 5. https://doi.org/10.1186/1756-0381-6-5.

Panicker, M.M., Minkley, E.G., 1992. Purification and properties of the F sex factor TraD protein, an inner membrane conjugal transfer protein. J. Biol. Chem. 267, 12761–12766.

Paton, J.C., Paton, A.W., 1998. Pathogenesis and diagnosis of Shiga toxin-producing *Escherichia coli* infections. Clin. Microbiol. Rev. 11, 450–479 (https://doi.org/file://Z:\References\Text Files\00000004469.txt).

Pielaat, A., Barker, G., Hendriksen, P., Hollman, P., Peijnenburg, A., Ter Kuile, B., 2013. A foresight study on emerging technologies: state of the art of Omics technologies and potential applications in food and feed safety. EFSA Support. Publ. 10. https://doi.org/10.2903/SP.EFSA.2013.EN-495.

Pielaat, A., Boer, M.P., Wijnands, L.M., van Hoek, A.H.A.M., Bouw, E., Barker, G.C., Teunis, P.F.M., Aarts, H.J.M., Franz, E., 2015. First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7 by coupling genomic data with in vitro adherence to human epithelial cells. Int. J. Food Microbiol. 213, 130–138. https://doi.org/10.1016/j.ijfoodmicro.2015.04.009.

Preußel, K., Höhle, M., Stark, K., Werber, D., 2013. Shiga toxin-producing *Escherichia coli* O157 is more likely to lead to hospitalization and death than non-O157 serogroups—except O104. PLoS One 8, e78180. https://doi.org/10.1371/journal.pone.0078180.

Ren, Y., Zhang, L., Suganthan, P.N., 2016. Ensemble classification and regression: recent developments, applications and future directions. IEEE Comput. Intell. Mag. 11, 41–53. https://doi.org/10.1109/MCI.2015.2471235.

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1–39. https://doi.org/10.1007/s10462-009-9124-7.

Sampson, B.A., Gotschlich, E.C., 1992. Elimination of the vitamin B12uptake or synthesis pathway does not diminish the virulence of *Escherichia coli* K1 or *Salmonella typhimurium* in three model systems. Infect. Immun. 60, 3518–3522.

Santerre, J., Boisvert, S., Davis, J., Xia, F., Stevens, R., 2015. Gene identification and strain classification using Random Forests. In: Great Lakes Bioinformatics Conference. Purdue University, West Lafayette, Indiana.

Schapire, R.E., 1990. The strength of weak learnability. Mach. Learn. 5, 197–227. https://doi.org/10.1023/A:1022648800760.

Seubert, A., Hiestand, R., De La Cruz, F., Dehio, C., 2003. A bacterial conjugation machinery recruited for pathogenesis. Mol. Microbiol. 49, 1253–1266. https://doi.org/10.1046/j.1365-2958.2003.03650.x.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C.T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R., 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat. Med. 8, 68–74. https://doi.org/10.1038/nm0102-68.

Spinale, J.M., Ruebner, R.L., Copelovitch, L., Kaplan, B.S., 2013. Long-term outcomes of Shiga toxin hemolytic uremic syndrome. Pediatr. Nephrol. 28, 2097–2105. https://doi.org/10.1007/s00467-012-2383-6.

Thorpe, C.M., 2004. Shiga toxin-producing *Escherichia coli* infection. Clin. Infect. Dis. 38, 1298–1303. https://doi.org/10.1086/383473.

Tian, C., Gregersen, P.K., Seldin, M.F., 2008. Accounting for ancestry: population substructure and genome-wide association studies. Hum. Mol. Genet. 17. https://doi.org/10.1093/hmg/ddn268.

Tobe, T., Beatson, S.A., Taniguchi, H., Abe, H., Bailey, C.M., Fivian, A., Younis, R., Matthews, S., Marches, O., Frankel, G., Hayashi, T., Pallen, M.J., 2006. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. Proc. Natl. Acad. Sci. 103, 14941–14946. https://doi.org/10.1073/pnas.0604891103.

Urbanowicz, R.J., Granizo-Mackenzie, A., Moore, J., 2012. An analysis pipeline with statistical and visualization-guided knowledge discovery for Michigan-style learning classifier systems. IEEE Comput. Intell. Mag. 7, 35–45. https://doi.org/10.1109/MCI.2012.2215124.

Van Melderen, L., De Bast, M.S., 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet. https://doi.org/10.1371/journal.pgen.1000437.

Velez, D.R., White, B.C., Motsinger, A.A., Bush, W.S., Ritchie, M.D., Williams, S.M., Moore, J.H., 2007. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. Genet. Epidemiol. 31, 306–315. https://doi.org/10.1002/gepi.20211.

Wallden, K., Rivera-Calzada, A., Waksman, G., 2010. Type IV secretion systems: versatility and diversity in function. Cell. Microbiol. https://doi.org/10.1111/j.1462-5822.2010.01499.x.

Whitney, D.H., Elashoff, M.R., Porta-Smith, K., Gower, A.C., Vachani, A., Ferguson, J.S., Silvestri, G.A., Brody, J.S., Lenburg, M.E., Spira, A., 2015. Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. BMC Med. Genet. 8, 18. https://doi.org/10.1186/s12920-015-0091-3.

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. https://doi.org/10.1201/b12207-2.