creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

A Thesis for the Degree of Doctor of Philosophy

Development of machine learning models to predict pathogenic potential and transcriptional regulatory network of enterohemorrhagic *Escherichia coli*

장출혈성대장균의 잠재적 독성 및 전사조절 네트워크 예측을 위한 기계학습 모델 개발

August 2022

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Hanhyeok Im

Development of machine learning models to predict pathogenic potential and transcriptional regulatory network of enterohemorrhagic *Escherichia coli*

장출혈성대장균의 잠재적 독성 및 전사조절 네트워크 예측을 위한 기계학습 모델 개발

지도교수 최 상 호

이 논문을 농학박사학위논문으로 제출함

2022 년 5 월

서울대학교 대학원

농생명공학부

임 한 혁

임한혁의 박사학위논문을 인준함

2022 년 7 월

위 원 장 _____(인)

부위원장 _____(인)

위 원 _____(인)

위 원 _____(인)

위 원 _____(인)

Abstract

Development of machine learning models to predict pathogenic potential and transcriptional regulatory network of enterohemorrhagic *Escherichia coli*

Hanhyeok Im

Department of Agricultural Biotechnology

The Graduate School

Seoul National University

Enterohemorrhagic *Escherichia coli* (EHEC) is a causative agent of human illnesses ranging from mild diarrhea to hemolytic uremic syndrome, which often results in permanent kidney failure. EHEC is considered as one of the major public health concerns because the novel pathogenic isolates continuously emerge and cause worldwide outbreaks. In order to control the disease burden of emerging EHEC, the accurate assessment of its pathogenic potential is critical. However, the conventional methods that use serotypes or several virulence genes have limitations in evaluating the emerging EHEC isolates with either unknown serotypes or a novel combination of virulence genes. In the present study, I developed a machine learning

(ML) model using the support vector machine (SVM) algorithm, the SVM model, to predict the pathogenic potential of the EHEC isolates using their whole genome sequencing data. The SVM model successfully predicted the pathogenicity of the isolates from the major sources of EHEC outbreaks, the isolates with a 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 of the SVM model, proposing the genes potentially essential for the pathogenicity of EHEC. Consequently, these results suggest that the SVM model is a more reliable and broadly applicable method to evaluate the pathogenic potential of EHEC isolates compared with conventional methods. Meanwhile, the elucidation of the transcriptional regulatory networks (TRNs) that are widely conserved in various EHEC isolates is a prerequisite for the prevention and treatment of the infection caused by newly emerging EHEC isolates. However, the analyses of current TRNs are still limited to comprehensively understand the target genes generally co-regulated under various conditions regardless of their genetic backgrounds. In this study, I developed a ML model using independent component analysis (ICA) algorithm, the ICA model, to decompose the large-scale transcriptome data of EHEC into the modulons, which contain the target genes of several TRNs. The locus of enterocyte effacement (LEE) and the Shiga toxin (Stx) modulons mainly consisted of the Ler regulon and the Stx prophage genes,

respectively, confirming that the ICA model properly grouped the co-regulated genes of EHEC. Further investigation revealed that the LEE modulon contained the Z0395 gene as a novel member of the Ler regulon, and the Stx modulon contained the *thi* and *cus* locus genes in addition to the Stx prophage genes. Concurrently, the Stx prophage genes were also regulated by thiamine and copper ions known to control the *thi* and *cus* locus genes, respectively. The modulons of EHEC effectively clustered the genes co-regulated regardless of the growth conditions and the genetic backgrounds. In addition, the changed activities of the individual modulons successfully explained the differential expressions of the virulence and survival genes during the course of infection of EHEC in bovine. Consequently, these results suggested that ICA of the large-scale transcriptome data can expand and enhance the current understanding of the TRNs of EHEC. Altogether, this study presents the ML models to analyze the large-scale genome and transcriptome data of EHEC and thereby investigate the pathogenic potential and TRNs of the pathogen. The ML-based approaches could be used to develop novel methods to prevent and control the infection caused by the newly emerging EHEC.

Keywords: EHEC, Machine learning, Genome, Pathogenic potential, Transcriptome, Transcriptional regulatory network

Student number: 2017-21443

Contents

List of Figures

List of Tables

Chapter I.

Background

I-1. Enterohemorrhagic *Escherichia coli*

Escherichia coli is a Gram-negative, facultative anaerobic, rod-shaped, and coliform bacterium, which belongs to the Enterobacteriaceae family. The bacteria are naturally found in the lower intestine of warm-blooded organisms. *E. coli* constitutes about 0.1% of human gut microbiota and has a mutualistic relationship with the host by providing essential nutrients, such as B vitamins, enhancing nutrient acquisition, and preventing the colonization of pathogenic bacteria in the intestine (Eckburg *et al.*, 2005; Gao *et al.*, 2014). Although most *E. coli* strains are harmless, some of them can cause serious foodborne illness in humans by acquiring virulence factors through plasmids, transposons, bacteriophages, and/or pathogenicity islands (Lim *et al.*, 2010). The pathogenic *E. coli* characterized by its ability to produce Shiga toxins are referred to as enterohaemorrhagic *E. coli* (EHEC) (World Health Organization *et al.*, 2018).

EHEC is a crucial cause of worldwide foodborne diseases. EHEC causes a wide range of human illnesses ranging from mild diarrhea to hemolytic uremic syndrome (HUS), which often results in end-stage renal disease with a high mortality rate (Smith and Fratamico, 2018). The World Health Organization (WHO) estimated that EHEC caused more than 1 million infection cases, resulting in more than 100 deaths and about 13,000 disability-adjusted life years (DALYs) worldwide (Havelaar *et al.*, 2015). In South Korea, EHEC infectious disease is designated as a group II National Notifiable Infectious Diseases. In the last 10 years from 2010 to 2019, 438 infection

cases have been reported by Korea Disease Control and Prevention Agency (Young-sun *et al.*, 2019). Recently, not only the outbreaks initiated by major serotype EHEC have been reported, but also the serotype of the causative isolates has been continuously changed (Young-sun *et al.*, 2019).

Although the order of top important sources of EHEC infection differs somewhat across worldwide regions, the meat and dairy products from domesticated ruminants, such as cattle, sheep, and goats, are recognized as the most important sources of EHEC, accounting for about 50% of total infection cases (World Health Organization *et al.*, 2018; Adams *et al.*, 2019; Koutsoumanis *et al.*, 2020). Among the products from ruminants, meat and dairy product from cattle are the most important food sources of EHEC in worldwide regions (Havelaar *et al.*, 2015; Hald *et al.*, 2016). The polluted water by livestock manure and consequent contamination of farm products are also major sources of EHEC, accounting for about 30% of total infection cases (World Health Organization *et al.*, 2018; Adams *et al.*, 2019; Koutsoumanis *et al.*, 2020). Overall, cattle, dairy products, and farm products are the most frequently identified sources of EHEC infection cases.

I-1-1. Virulence factors of EHEC

Shiga toxin (Stx)

The clinical symptoms of the EHEC infections such as bloody diarrhea and HUS are induced by Stx, one of the major virulence factors of EHEC (Spinale *et al.*,

2013). The Stx is composed of two major subunits A and B. The A subunit binds noncovalently to a pentamer of five identical B subunits. The pentamer of B subunits binds to a component of the cell membrane known as glycolipid globotriaosylceramide (Gb3) and derives the toxin-mediated internalization (Römer *et al.*, 2007). The A subunit of internalized toxin not only inhibits protein synthesis but also triggers many signaling cascades that influence cytokine secretion and induce cell death by apoptosis (Johannes and Römer, 2010). The Stx is effective against small blood vessels found in the digestive tract or kidney. Especially, the toxin appears to be highly harmful to the glomerulus, the filtering structure of the kidney, because the vascular endothelium of the structure expresses high levels of Gb3 on the cell surface (Obrig, T. G. *et al.*, 1993). Disruption of these structures by Stx leads to end-stage renal disease that has high mortality rates (Young-sun *et al.*, 2019). Although the Stx is an essential virulence factor to cause HUS, not all Stx is associated with the disease. Stx has 2 major types (Stx1 and Stx2). The Stx1 has four subtypes (a, c, d, and e) and the Stx2 has 12 subtypes (a to l) (European Food Safety Authority, 2013; Koutsoumanis *et al.*, 2020). Among the subtypes, Stx1a, Stx2a, Stx2c, and Stx2d are most consistently associated with HUS. However, even the association is not as definitive nor conclusive (European Food Safety Authority, 2013; Koutsoumanis *et al.*, 2020).

Since the Stx is encoded in the genome of lambdoid bacteriophages, the Stx genes are expressed by the activation of the phage lytic cycle (Johansen *et al.*, 2001). The

expression of lambdoid bacteriophage genes is controlled by antitermination whereby the early promoters, P_L and P_R, and late promoter, $P_{R'}$, are terminated at downstream intrinsic terminators, t_L, t_R, and $t_{R'}$, respectively. The lytic cycle is induced by the RecA-mediated SOS response and repressions of P_L and P_R are relieved, leading to the expression of antiterminators N and Q. The antiterminators prevent termination at t_L, t_R, and $t_{R'}$, leading to the expression of phages genes. Especially, antiterminator Q prevents the termination by $t_{R'}$, allowing the late $P_{R'}$ transcript including the Stx (Pacheco and Sperandio, 2012; Sy *et al.*, 2020). Thus, the expression of Stxs is related to that of antiterminator Q.

Locus of enterocyte effacement (LEE)

Production of Stx alone without adherence is regarded as insufficient for EHEC to cause severe infections (World Health Organization, 2018). Therefore, the ability to adhere to intestinal epithelial cells is a critical property in determining the key virulence traits of EHEC. The LEE is one of the major virulence factors of EHEC, mediating the attachment between the pathogen and host cell (World Health Organization *et al.*, 2018; Koutsoumanis *et al.*, 2020). The LEE contains the genes encoding the type III secretion system (T3SS) and adherence proteins. These virulence factors are necessary to form attaching and effacing (A/E) lesions, the pedestal-like attachment structure characterized by the accumulation of polymerized actin (Sheng *et al.*, 2006; Abu-Ali *et al.*, 2009).

LEE consists of five open reading frames (LEE1-5), and its regulation is mainly controlled by the LEE master regulator, Ler encoded by *ler* (Platenkamp and Mellies, 2018). Ler is encoded in the LEE1 and is a member of the H-NS family (Torres *et al.*, 2007). Although all other members of the H-NS family transcriptional factors (TFs) known to date act as a repressor, Ler is an activator of LEE2-5 (Torres *et al.*, 2007; Platenkamp and Mellies, 2018). LEE2-5 encodes an adhesin, intimin, and its receptor, Tir, responsible for intimate attachment, several secreted proteins, and their chaperones. The secreted proteins consist of effectors as well as translocators (EspA, EspD, and EspB) required for translocating effectors into host cells. Five LEE-encoded effectors (Tir, EspG, EspF, Map, and EspH) have been identified, which are involved in modulating the host cytoskeleton (Torres *et al.*, 2007; Platenkamp and Mellies, 2018). Ler also activates the non-LEE located genes, such as *nleI/G* to *nleF*, *etp* operon genes and *stcE*, and *lpxR* (Grys *et al.*, 2005; Abe *et al.*, 2008; Ogawa *et al.*, 2018). The non-LEE encoded effector proteins encoded by *nle* genes are known to be required for colonization in the host colon (Abe *et al.*, 2008). The metalloprotease StcE and its Etp type II secretion system (T2SS) encoded by *stcE* and *etp* operon genes, respectively, contribute to the intimate adherence of EHEC by cleavaging the mucin layers of the host cell (Grys *et al.*, 2005). The lipid A 3′-O-deacylase encoded by *lpxR* reduces the inflammatory response by remodeling the lipid A structure (Ogawa *et al.*, 2018).

pAA plasmid

Highly pathogenic EHEC commonly carries LEE as an adherence factor, but a small subset of the pathogen carries pAA plasmid. The plasmid includes various virulence genes encoding aggregative adherence fimbriae (AAF), dispersin, and type I secretion system (T1SS) (Prieto *et al.*, 2021). AAFs encoded by *aaf* genes promote agglutination of planktonic cells, adherence to the intestinal epithelium, and formation of biofilm for persistent colonization (Okhuysen and DuPont, 2010; Jønsson *et al.*, 2017). Dispersin and its T1SS enable the pathogen to disperse across the intestinal epithelial surface (Sheikh *et al.*, 2002).

The virulence genes included in the pAA plasmid are positively regulated by AggR encoded by *aggR*, which is a member of the AraC-XylS family (Prieto *et al.*, 2021). AggR autoactivates its own transcription (Morin *et al.*, 2010). The expression of *aggR* is also activated or repressed by the factor for inversion stimulation (Fis) and H-NS global regulators, respectively (Morin *et al.*, 2010). In addition, the *aggR* is induced according to the level of guanosine 5'-diphosphate 3'-diphosphate (ppGpp) or guanosine 5'-triphosphate 3'-diphosphate (pppGpp), (p)ppGpp, which are synthesized by the bacteria in response to the nutrient-limited conditions.

I-1-2. Assessment of EHEC pathogenic potential

Since the huge outbreak was caused by EHEC serotype O157:H7 in 1982, the serotype has been used as a criterion to assess the pathogenic potential of the isolates

that cause severe human diseases (Levine, 1987; World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Based on the criterion, the EHEC isolates are categorized into the O157 serotype or the non-O157 serotype EHEC. The isolates with the O157 serotype are characterized by usually possessing *stx2a* and LEE genes (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Among the non-O157 isolates, the isolates with O26, O157, O121, O145, O111, O104, O91, O103, and O55 serotypes are recognized as pathogenic (European Food Safety Authority, 2013; Gould *et al.*, 2013; Eichhorn *et al.*, 2015; World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Especially, the isolates with the O104 serotype are characterized by possessing the *stx2a* and *aggR*. However, the isolates with the same serotype may have different pathogenic potential because many EHEC virulence genes are mobile and thus can be lost or transferred to other bacteria. As a result, serotype alone is not reliable for evaluating the pathogenic potential of EHEC isolates (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). The potential risk of EHEC isolates causing severe illness is best predicted by using the existing knowledge of EHEC virulence factors such as Stx, LEE, and pAA. Nevertheless, the minimum virulence gene combinations anticipated to cause severe illness is still unknown due to the high genome diversity of EHEC. The genome diversity also makes it difficult to comprehensively understand the molecular mechanisms of the EHEC pathogenicity (World Health Organization, 2018).

I-2. Machine learning

Machine learning (ML) is a program that uses specific computer algorithms that can use data to automatically improve performance for a specific task, such as predictions or classification, without being explicitly programmed to do so (Koza *et al.*, 1996). ML is becoming more popular because it has notable performance in fields of big and complex data where traditional statistical methods are difficult to gain valuable information (Eisenstein and Dodd, 1982; Houle *et al.*, 2010; Yang *et al.*, 2015; Bzdok *et al.*, 2018). In addition, ML-based analyses are becoming more accessible to researchers because the computational facilities and ML algorithms are increasingly and widely available.

In the field of biology, advances in sequencing technology have significantly increased the submission of sequencing data to public databases (O'Leary *et al.*, 2016). The public databases have stored massive amounts of sequencing data of genome and transcriptome for various organisms, enabling large-scale bioinformatic analyses combined with ML. For example, the ML-based bioinformatic analyses have been conducted to predict the bacteria transcriptional regulatory network (Sastry *et al.*, 2019), antibiotic resistance (Moradigaravand *et al.*, 2018), host adaptability (Lupolova *et al.*, 2017), zoonotic or pathogenic potential (Lupolova *et al.*, 2016, 2019).

I-2-1. Unsupervised ML algorithms

The ML algorithms include two broad categories: unsupervised and supervised. The unsupervised ML algorithms identify inherent patterns in given data without the concept of output and then discriminate the data using the inherent patterns (Svensson *et al.*, 2014; Lupolova *et al.*, 2019). The phylogenetic tree analysis, principal component analysis (PCA), independent component analysis (ICA), and Gaussian mixture model (GMM) are examples of the analyses using unsupervised ML algorithms. The phylogenetic tree analysis visualizes the branching diagram that represents the pairwise distances between the features of given data, showing the relative relationships and structural clustering of the data (Baum *et al.*, 2005). PCA computes principal components, which are the unit vectors that best describe the given data (Pearson, 1901). ICA also calculates vectors for a given data, but unlike PCA, it finds the vectors of maximum independence rather than unit vectors orthogonal to each other (Hyvärinen and Oja, 2000). GMM identifies the distinct Gaussian distributions with different parameters to discriminate the given data (Guoshen Yu *et al.*, 2012).

I-2-2. Supervised ML algorithms

Unlike the unsupervised ML algorithms without the concept of output, the supervised ML algorithms, such as Gaussian naive Bayes (GaussianNB), decision trees (DTs), random forest (RF), support vector machine (SVM), and multi-layer

perceptron (MLP), predict an output from an input data. Thus, the supervised ML algorithms need to be trained on known input-output pairs, also known as training dataset, until they can predict the correct output using the given input data (Lupolova *et al.*, 2019). GaussianNB improves its prediction performance by learning the relationship between the features of input data and output based on the assumption that the features are independent of each other based on the Bayesian principle (John and Langley, 2013). DTs predict the output by learning simple if-then functions, also known as decision rules, inferred from the training dataset (Loh, 2011). RF is also a decision rule-based ML algorithm that leverages the prediction performance by combining multiple DT models (Breiman, 2020). SVM predicts the output based on the hyperplane which is determined by itself to most correctly classify the training dataset in high-dimensional space (Cortes and Vapnik, 1995). Finally, MLP predicts the output by using the trained multi-layers of perceptron that imitate the biological neural networks of the brain (Collobert and Bengio, 2004).

I-3. Objective of this study

EHEC causes a wide range of human illnesses ranging from mild diarrhea to HUS, which often results in permanent kidney failure. In order to prevent and control the infection cases and outbreaks caused by EHEC, it is required to evaluate the pathogenicity of EHEC and understand its molecular mechanisms. However, the conventional serotyping and several virulence gene combination methods have limitations in estimating the pathogenic potential of emerging EHEC isolates with non-O157 serotypes which carry novel virulence genes combination. Meanwhile, the identification of the widely conserved transcriptional regulatory networks (TRNs) of various EHEC isolates is required for the prevention and treatment of the emerging EHEC infection. Nevertheless, studies for the current EHEC TRNs are still limited in comprehensively understanding their target genes generally co-regulated under various conditions regardless of the genetic backgrounds. In this study, I built ML models that predict the pathogenic potential and transcriptional regulatory networks (TRNs) of EHEC. The ML model to predict the pathogenic potential of EHEC was designed to use whole genome sequencing data rather than the presence/absence of several virulence genes. The ML model to predict the TRNs of EHEC was designed to use large-scale transcriptome data to identify the inherently co-regulated genes under various conditions regardless of the genetic backgrounds of EHEC.

Chapter II.

Pathogenic potential assessment of the enterohemorrhagic *Escherichia coli* by a source attribution-considered machine learning model

Part of this work in Chapter II was published in *Proceedings of the National Academy of Sciences* in 2021, as an article entitled "Pathogenic potential assessment of the Shiga toxin-producing *Escherichia coli* by a source attribution-considered machine learning model".

II-1. Introduction

Emerging pathogens causing an increasing number of outbreaks are now considered as a major risk to public health (Vouga and Greub, 2016). The exact assessment of the pathogenic potential of pathogens is required to predict and manage their health risk in advance (Smith and Fratamico, 2018). Conventional methods such as serotyping and virulence gene combinations have been used to assess bacterial pathogenic potentials (Tauxe, 2002; Koutsoumanis *et al.*, 2020). 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 (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Therefore, the development of novel methods assessing their pathogenic potential is required to cope with the public health risks caused by newly emerging pathogens.

Enterohemorrhagic *Escherichia coli* (EHEC) causes a wide range of human illnesses ranging from mild diarrhea to hemolytic uremic syndrome, which often results in permanent kidney failure (Karmali, 2017). In addition to the O157 serotype EHEC, emerging non-O157 serotype EHECs have been identified as causative agents for the increasing outbreaks lately (European Food Safety Authority, 2013; World Health Organization, 2018). 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 EHECs has limitations (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). It has been reported that virulence genes such as *stx2* and *eae* are required for the pathogenesis of EHEC (European Food Safety Authority, 2013; Fratamico *et al.*, 2016; Naseer *et al.*, 2017; World Health Organization, 2018; Koutsoumanis *et al.*, 2020). However, the emerging highly pathogenic EHEC isolates carry novel virulence genes (World Health Organization, 2018; Koutsoumanis *et al.*, 2020), and indeed, an EHEC isolate with a novel combination of *stx2* and *aggR* had caused a huge outbreak in Europe, 2011 (European Food Safety Authority, 2013; Boisen *et al.*, 2015).

Recently, advances in next-generation sequencing technologies have enabled us to exploit whole genome sequencing (WGS) data (Houle *et al.*, 2010; Franz *et al.*, 2014). 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 into their pathogenicity by using traditional statistical methods (Houle *et al.*, 2010; Yang *et al.*, 2015). In contrast, machine learning (ML) algorithms have notable performance in the analysis of the complex WGS data (Houle *et al.*, 2010; Yang *et al.*, 2015) and thus have been exploited lately to find out the connection between genetic features and pathogenicity of some pathogens (Houle *et al.*, 2010; Lupolova *et al.*, 2016, 2017, 2019; Moradigaravand *et al.*, 2018). 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 (Svensson *et al.*, 2014; Lupolova *et al.*, 2019). On the other hand, the supervised ML algorithms such as Gaussian naive Bayes (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 (Lupolova *et al.*, 2019).

In this study, I built various ML models and compared their performances in evaluating the pathogenicity of EHEC isolates. I subsequently developed an ML model using the SVM algorithm, the SVM model, which can evaluate the pathogenic potential of the EHEC isolates most accurately among the tested ML models. Because the SVM model can also estimate the pathogenic potential of EHEC isolates of which the pathogenicity cannot be estimated by conventional methods, the model is more widely applicable to predict the risk of EHEC 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 EHEC isolates.

II-2. Materials and Methods

II-2-1. Generation of the input dataset for ML

The WGS data and metadata of 3,303 EHEC isolates were retrieved from the GenBank database at the NCBI (https://www.ncbi.nlm.nih.gov/genbank/) (Table II-1). The quality of the WGS data was checked using Kraken2, a taxonomy classification tool (Wood *et al.*, 2019), and QUAST, a quality assessment tool for genome assemblies (Gurevich *et al.*, 2013). The quality-passed WGS data were annotated using Prokka, a prokaryotic genome annotation program (Seemann, 2014). The classification of the clinical and environmental isolates was determined based on the metadata of the isolates. The 2,292 clinical EHEC isolates were set as the positive control group (pathogenic). The EHEC isolates from the cattle, dairy products, and farm products have been reported as the major sources of outbreaks, which may have high pathogenic potential (World Health Organization, 2011, 2018; Koutsoumanis *et al.*, 2020). 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 EHEC isolates was constructed using PIRATE, a pangenomics toolbox (Bayliss *et al.*, 2019). The genes statistically relevant to either positive or negative control group were selected as significant genes by Scoary, a pangenome-wide association studies (pan-GWAS)

tool ($p < 0.05$) (Brynildsrud *et al.*, 2016), 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) (Liu *et al.*, 2019; UniProt Consortium and Bateman, 2019). The subtypes of the Shiga toxin were identified using the reference sequences of the Shiga toxin subtypes (Scheutz *et al.*, 2012). Sequence alignments with the reference sequences were conducted by using DIAMOND, a sequence alignment tool (Buchfink *et al.*, 2015). The presence/absence matrix of the significant genes and Shiga toxin subtype genes was used as the input dataset and can be found in Table II-2.

Table II-1. The WGS data and metadata of EHEC isolates

NCBI accession number	Clinical/ environmental	Assembly level	Host	isolation source	Classification
GCA_000464955.2	clinical	complete	Homo sapiens	stool	clinical
GCA_000520035.1	clinical	complete	Homo sapiens	feces; clinical sample	clinical
GCA_000520055.1	clinical	complete	Homo sapiens	feces	clinical
GCA_000967155.1	clinical	complete	Homo sapiens	NULL	clinical
GCA_000986765.1	clinical	complete	Homo sapiens	stool sample	clinical
GCA_001420935.1	clinical	complete	Homo sapiens	NULL	clinical
GCA_001420955.1	clinical	complete	Homo sapiens	stool	clinical
GCA_001644725.1	clinical	complete	Homo sapiens	stool	clinical
GCA_001644745.1	clinical	complete	Homo sapiens	stool	clinical
GCA_001645235.2	clinical	complete	Homo sapiens	stool	clinical
GCA_001677515.1	clinical	complete	Homo sapiens	bagged lettuce	clinical
GCA_001721125.1	clinical	complete	Homo sapiens	stool	clinical
GCA_001721225.1	clinical	complete	Homo sapiens	NULL	clinical
GCA_001890205.1	clinical	complete	NULL	human	clinical
GCA_001890225.1	clinical	complete	NULL	human	clinical
GCA_001890245.1	clinical	complete	NULL	human	clinical
GCA_001890265.1	clinical	complete	NULL	human	clinical
GCA_001890325.1	clinical	complete	NULL	human	clinical
GCA_002208865.2	clinical	complete	Homo sapiens	clinical isolate	clinical
GCA_002214745.2	clinical	complete	Homo sapiens	human stool	clinical
. . .	. . .	. . .	. . .	. . .	. . .

Data from only partial isolates are presented because the entire data of whole isolates are too large to be displayed in the table.

Full data can be found at Supplementary Information (SI) Appendix Dataset S1 (https://doi.org/10.1073/pnas.2018877118).

Table II-2. The presence/absence matrix of the significant genes and Shiga toxin subtype genes

Accession number	Gene presence or absence[a] / clinical or environmental classification								
	g011091	g001329_1	g001062_2	· · ·	*stx2d*	stx2e	stx2f	stx2g	classification[b]
GCA_000464955.2	0	0	0	· · ·	0	0	0	0	1
GCA_000520035.1	0	0	1	· · ·	0	0	0	0	1
GCA_000520055.1	1	0	1	· · ·	0	0	0	0	1
GCA_000662395.1	1	0	1	· · ·	0	0	0	0	0
GCA_000967155.1	1	0	1	· · ·	0	0	0	0	1
GCA_000986765.1	1	0	1	· · ·	0	0	0	0	1
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
GCA_008757105.1	0	1	1	· · ·	0	0	0	0	1
GCA_008757195.1	1	0	1	· · ·	0	0	0	0	1
GCA_008757255.1	1	0	0	· · ·	0	0	0	0	1
GCA_008757265.1	1	1	1	· · ·	0	0	0	0	1
GCA_008757335.1	1	1	1	· · ·	0	0	0	0	1
GCA_008757345.1	1	1	1	· · ·	0	0	0	0	1
GCA_008757375.1	1	1	1	· · ·	0	0	0	0	1

[a] 1 or 0 indicate the presence or the absence of genes tagged by EHEC pangenome.

[b] 1 or 0 indicate the clinical or environmental EHEC isolates, respectively.

Data from only partial isolates are presented because the entire data of whole isolates are too large to be displayed in the table.

Full data can be found at SI Appendix Dataset S2 (https://doi.org/10.1073/pnas.2018877118).

II-2-2. 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 (Stamatakis, 2014). The reliability of internal branches was assessed by bootstrapping based on 500 replicates. The phylogenetic tree was visualized by the *ggtree* R package (Yu *et al.*, 2017). PCA for the input dataset was conducted with the Scikit-learn python package (Varoquaux *et al.*, 2015). The GMM models were built using the Scikit-learn python package (Varoquaux *et al.*, 2015) and trained on the PCA transformed input dataset. The Matplolib python package was used to visualize the PCA and GMM model results (Hunter, 2007).

II-2-3. Supervised ML: GaussianNB, DTs, RF, and SVM

Four different supervised ML algorithms, GaussianNB (Jahromi and Taheri, 2018), DTs (Kotsiantis, 2013), RF (Belgiu and Drăguţ, 2016), and SVM (Bhavsar and Panchal, 2012) 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. The grid searches were conducted for the three supervised ML models: decision trees (DTs), random forest (RF), and support vector machine (SVM) model to find out optimized hyperparameters of the models. The

optimized hyperparameters of the DTs model were the following: criterion = entropy, splitter = best, max_depth = 49, min_samples_leaf = 1, min_samples_split = 5, class_weight = balanced. The optimized hyperparameters of the RF model were the following: criterion = gini, max_depth = 20, min_samples_leaf = 10, min_samples_split = 5, class_weight = balanced, n_estimators = 1,800. The optimized hyperparameters of the SVM model were the following: kernel = rbf, gamma = 0.0001, C = 10, class_weight = balanced. The grid searches and construction of the models were conducted using a Scikit-learn python package (Varoquaux *et al.*, 2015). The construction and optimization of a single hidden layer multilayer perceptron (MLP) model were conducted by using a TensorFlow python package (Abadi *et al.*, 2016). The optimized hyperparameters of the MLP model were the following: the unit number of a hidden layer = 3,000, dropout rate = 0.2, optimizer = RMSProp, learning rate = 0.0001, and batch size = 100. The number of epochs to train the model was set as 500 with an early stopping condition. 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 Matthews correlation coefficient (MCC) and the area under the receiver operating characteristic curve (AUROC) scoring methods were further used

to produce a more informative and truthful score (Berrar and Flach, 2012; Chicco and Jurman, 2020). 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 (Varoquaux *et al.*, 2015). The bar plots of the scores and the ROC curve plots were visualized by the Seaborn python packages (https://seaborn.pydata.org/).

II-2-4. Examination of the SVM model using the decision function values

The SVM model calculates decision function values to classify each EHEC 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 (Platt, 1999). 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/), and their decision function values were plotted as swarm plots. The serotypes of the EHEC isolates were identified with the SerotypeFinder CGE tool (Joensen *et al.*, 2015). The

conventional virulence gene combination method used the combinations of *stx2a* or *stx2d* with an additional adherence factor *eae* or *aggR* (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Thus, the isolates were grouped by the following combinations: *stx2a*, *stx2a* + *eae* or *aggR*, and *stx2d* (*stx2d* + *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/).

Table II-3. The metadata of the EHEC isolates with a history of outbreak

NCBI accession number	BioProject number	Assembly level	Outbreak	Classification
GCA_000022225.1	PRJNA30045	Complete	2006 USA spinach	environmental/other
GCA_000181755.1	PRJNA27733	Scaffold	2007 USA spinach	environmental/other
GCA_000350045.2	PRJNA73635	Contigs	2010 Sweden cases	environmental/other
GCA_000235245.1	PRJNA68211	Scaffold	2011 Europe	environmental/other
GCA_000235225.1	PRJNA68213	Scaffold	2011 Europe	environmental/other
GCA_000235205.1	PRJNA68215	Scaffold	2011 Europe	environmental/other
GCA_000235185.1	PRJNA70733	Scaffold	2011 Europe	environmental/other
GCA_000235165.1	PRJNA70735	Scaffold	2011 Europe	environmental/other
GCA_000235145.1	PRJNA70737	Scaffold	2011 Europe	environmental/other
GCA_000235125.1	PRJNA70739	Scaffold	2011 Europe	environmental/other
GCA_000235105.1	PRJNA70741	Scaffold	2011 Europe	environmental/other
GCA_000235085.1	PRJNA70743	Scaffold	2011 Europe	environmental/other
GCA_000235065.1	PRJNA70745	Scaffold	2011 Europe	environmental/other
GCA_000235045.1	PRJNA70747	Scaffold	2011 Europe	environmental/other
GCA_000217975.2	PRJNA67929	Scaffold	2011 UK cases	environmental/other
GCA_010915745.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010915715.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010915835.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other

GCA_010915695.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820905.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820875.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820785.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820795.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820755.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820745.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820725.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820675.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820865.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_010820825.1	PRJNA517910	Contig	2018 USA romaine lettuce	environmental/other
GCA_900000205.1	PRJEB7864	Contig	France cow's cheese	environmental/other
GCA_012708905.1	PRJNA218110	Contig	UK cattle	clinical
GCA_013167715.1	PRJNA528413	Complete	UK cattle	clinical
GCA_013167695.1	PRJNA528413	Complete	UK cattle	clinical
GCA_013167675.1	PRJNA528413	Complete	UK cattle	clinical
GCA_000513035.1	PRJNA63279	Contig	UK cattle	clinical
GCA_013167635.1	PRJNA528413	Complete	UK cattle	clinical
GCA_013167615.1	PRJNA528413	Complete	UK cattle	clinical
GCA_013167595.1	PRJNA528413	Complete	UK cattle	clinical
GCA_013167575.1	PRJNA528413	Complete	UK cattle	clinical
GCA_012707985.1	PRJNA218110	Contig	UK cattle	clinical

GCA_012708165.1	PRJNA218110	Contig	UK cattle	clinical
GCA_012708205.1	PRJNA218110	Contig	UK cattle	clinical
GCA_013168175.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013168155.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013168135.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013168055.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013343595.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167475.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167455.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167435.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167315.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167275.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167235.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_013167175.1	PRJNA528413	Complete	UK cattle	environmental/other
GCA_000471485.1	PRJNA215830	Scaffold	USA Saint Louis cases	environmental/other
GCA_000471505.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471525.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000466625.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471225.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471245.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471265.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471285.1	PRJNA215830	Scaffold	USA salad bar	environmental/other

GCA_000471305.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471325.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471345.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471365.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471385.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471405.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471065.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471425.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471445.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471465.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471085.2	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471105.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471125.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471145.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471165.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471185.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471205.1	PRJNA215830	Scaffold	USA salad bar	environmental/other
GCA_000471585.1	PRJNA215830	Scaffold	USA Washington cases	environmental/other
GCA_000471605.1	PRJNA215830	Scaffold	USA Washington cases	environmental/other
GCA_000471545.1	PRJNA215830	Scaffold	USA Webster County cases	environmental/other
GCA_000471565.1	PRJNA215830	Scaffold	USA Webster County cases	environmental/other

II-2-5. 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 (Altmann *et al.*, 2010). The analysis, however, tends to underestimate the importance of the genes which are highly correlated to others (Toloşi and Lengauer, 2011). Thus, gene clusters based on the Spearman rank-order correlation were generated and the input dataset gene clusters were used 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 (Varoquaux *et al.*, 2015). Functional categories of the genes were assigned based on the clusters of orthologous groups (COG) proteins database with the eggNOG-mapper, a functional annotation tool (Huerta-Cepas *et al.*, 2017). 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 Table II-4**.** PHAge Search Tool – Enhanced Release (PHASTER) was used to identify the phage genes in the input dataset (Arndt et al., 2016)

Table II-4. The weight values and the functional categories of the input dataset genes

Gene tag	Average weight value	Std	UniProt gene name	UniProt annotation
g002124_1	0.00401	0.00194		Antitermination protein Q
g007884	0.00393	0.00319		hypothetical protein
g003829	0.00392	0.00073		hypothetical protein
g005099	0.00392	0.00073		hypothetical protein
g004264	0.00392	0.00073		hypothetical protein
g004621	0.00392	0.00073		hypothetical protein
g006123	0.00366	0.00220		hypothetical protein
g000997	0.00346	0.00216	*espF(U)*	Secreted effector protein EF(U)
g006049	0.00331	0.00176		hypothetical protein
g011033	0.00317	0.00163		hypothetical protein
g001239_1	0.00316	0.00195		hypothetical protein
g013010	0.00304	0.00072	*lsoA*	mRNA endoribonuclease LsoA
g003934	0.00289	0.00166	*quuQ*	Prophage antitermination protein Q homolog QuuQ
g008365	0.00287	0.00045		hypothetical protein
g002431_1	0.00287	0.00045	*yaiX*	Putative uncharacterized acetyltransferase YaiX
g006097_1	0.00272	0.00141		hypothetical protein
g008512_1	0.00266	0.00205		hypothetical protein
g000970_1	0.00257	0.00101	*ydeR*	Uncharacterized fimbrial-like protein YdeR
g000812_2	0.00257	0.00069		hypothetical protein
g004651	0.00257	0.00069		hypothetical protein
g000779	0.00249	0.00442	*tufI*	Elongation factor Tu 1
g010304	0.00245	0.00159		hypothetical protein
g008258	0.00242	0.00074		SAR-endolysin
g000780_1	0.00242	0.00074	*bet*	Recombination protein bet
g000297_1	0.00242	0.00074	*yagK*	Uncharacterized protein YagK

g001401_1	0.00241	0.00180		hypothetical protein
g002721_1	0.00240	0.00100		hypothetical protein
g001055_1	0.00234	0.00184		hypothetical protein
g000139_2	0.00226	0.00075		hypothetical protein
g004263_1	0.00226	0.00101	*ant*	Antirepressor protein ant
g000276_1	0.00226	0.00121		hypothetical protein
g001348_2	0.00211	0.00100	*mrr*	Mrr restriction system protein
g000258_1	0.00207	0.00185	*ydfD*	Uncharacterized protein YdfD
g007977_1	0.00199	0.00143		hypothetical protein
g000279_3	0.00199	0.00141	*yoaE*	UPF0053 inner membrane protein YoaE
g017528	0.00199	0.00062		hypothetical protein
g000315	0.00190	0.00223		hypothetical protein
g000423	0.00188	0.00154		Regulatory protein CII
g007397	0.00185	0.00045		hypothetical protein
g019707	0.00185	0.00045		hypothetical protein
g000368	0.00181	0.00221	*ant*	Antirepressor protein ant
g018892	0.00168	0.00051		hypothetical protein
g006731	0.00166	0.00072		hypothetical protein
g000962_1	0.00166	0.00125	*kilR*	Killing protein KilR
g000215	0.00151	0.00000	*perC*	Protein PerC
g005674_1	0.00151	0.00000		hypothetical protein
g000382_1	0.00151	0.00000	*ydaV*	Uncharacterized protein YdaV
g013212	0.00151	0.00000		hypothetical protein
g008812	0.00151	0.00000		hypothetical protein
. . .	. . .	. . .	. . .	. . .

Data from only partial genes are presented because the entire data of whole pangenome genes are too large to be displayed in the table.

Full data can be found at SI Appendix Dataset S4 (https://doi.org/10.1073/pnas.2018877118).

II-2-6. 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.

II-3. Results

II-3-1. 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 EHEC 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 (II-2-1 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 pan-GWAS (Brynildsrud *et al.*, 2016). As a result, a total of 3,453 significant genes, including 148 virulence genes, were selected (Fig. II-1). The 148 virulence genes included the major virulence genes of the pathogenic EHEC, such as *eae*, *aggR*, and the locus of enterocyte effacement (LEE) effector protein genes (Table II-5) (Kaper *et al.*, 2004; Pacheco and Sperandio, 2012; Boisen *et al.*, 2015). 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 (Fig. II-1). 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 non-accidental. It has been reported that the subtypes of Shiga toxins are also associated with the pathogenicity of EHEC (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). Thus, the 10 Shiga toxin genes, *stx1a*, *stx1c*, *stx1d*, *stx2a*, *stx2b*, *stx2c*, *stx2d*, *stx2e*, *stx2f*, and *stx2g*, were added to the 3,453 significant genes. Accordingly, the presence/absence matrix of the 3,463 genes of the 2,646 EHEC isolates was used as an input dataset of the ML models for further analysis.

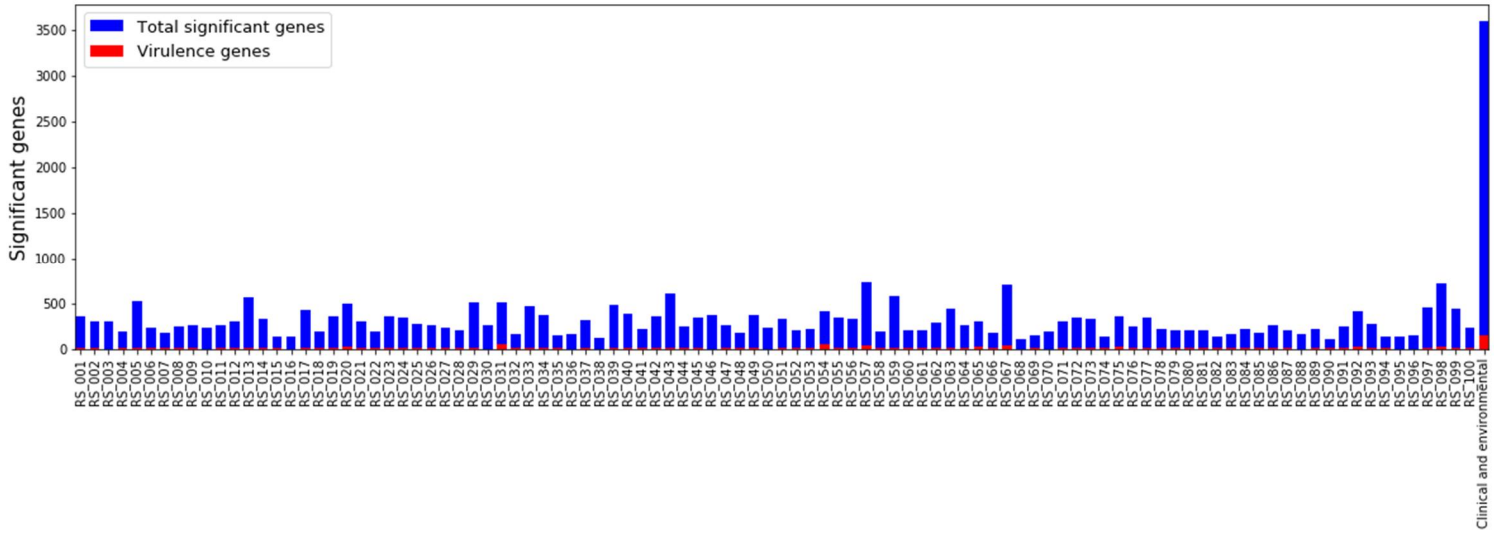

Figure II-1. The bar plots of the number of significant genes selected by pan-GWAS. When the clinical and environmental isolates were set as the positive and negative control group, respectively, the pan-GWAS result is labeled with 'Clinical and environmental.' As a result, a total of 3,453 significant genes, including 148 virulence genes, were selected. The panGWAS results of the trial groups are labeled with the random seed numbers of replicates (RS_number on the bottom). The average numbers of selected significant genes and virulence genes of the trial groups were 285.9 ± 131.4 and 9.3 ± 10.5, respectively. Error represents the SD.

Table II-5. The virulence genes selected by pan-GWAS using the clinical and environmental isolates as positive and negative control groups.

Class of virulence genes	Gene tag[a]	Gene name[b]	Virulence factor[c]
Adherence	g002621	*afaA*	Afimbrial adhesin
	g004539	*csgA*	Curli fibers
	g006896	*aap*	Dispersin
	g003533	*ecpA*	*E. coli* common pilus
	g002223	*elfA*	*E. coli* laminin-binding fimbriae
	g000169_3	*elfC*	*E. coli* laminin-binding fimbriae
	g001884	*elfG*	*E. coli* laminin-binding fimbriae
	g001806	*eae*	Intimin
	g003171	*faeJ*	K88 fimbriae
	g000521	*papA*	P fimbriae
	g001225_1	*papB*	P fimbriae
	g005728	*paa*	Porcine attaching-effacing associated protein
	g001225_2	*sfaB*	S fimbriae
	g000868_1	*toxB*	Adherence factor
	g000169_2	*fimD*	Type I fimbriae
	g004030	*fimF*	Type I fimbriae
	g000970_2	*fimG*	Type I fimbriae
	g002518	*fimH*	Type I fimbriae
Autotransporter	g004854	*aatA*	Adhesin involved in diffuse adherence
	g000470_4	*tibA*	Adhesin involved in diffuse adherence
	g000100_1	*agn43*	Adhesin involved in diffuse adherence
	g000129_1	*cdiA*	Contact-dependent inhibition
	g000869	*cdiB*	Contact-dependent inhibition
	g000194_2	*ehaA*	Contact-dependent inhibition
	g000115_3	*air/eaeX*	Enteroaggregative immunoglobulin repeat protein
	g000161_1	*espP*	Extracellular serine protease
	g000146_2	*upaH*	Adhesin involved in diffuse adherence
Invasion	g000403_1	*tia*	Adhesin and invasin
Iron uptake	g004196_1	*iutA*	Aerobactin siderophore
	g004819	*iucD*	Aerobactin siderophore
	g005463	*iucC*	Aerobactin siderophore
	g005647	*iucB*	Aerobactin siderophore
	g004753	*iucA*	Aerobactin siderophore
	g004919	*chuS*	Hemin uptake
	g002699	*chuA*	Hemin uptake
	g004964	*chuT*	Hemin uptake
	g005511	*chuW*	Hemin uptake

	g007941	*chuX*	Hemin uptake
	g005804	*chuY*	Hemin uptake
	g005629	*chuU*	Hemin uptake
	g001554	*ireA*	Iron-regulated element
	g004884	*iroB*	Salmochelin siderophore
	g004276	*ybtS*	Yersiniabactin siderophore
	g003294_1	*ybtX*	Yersiniabactin siderophore
	g003294_2	*ybtQ*	Yersiniabactin siderophore
	g005458	*ybtP*	Yersiniabactin siderophore
	g007459	*ybtA*	Yersiniabactin siderophore
	g001015	*irp2*	Yersiniabactin siderophore
	g000825_1	*irp1*	Yersiniabactin siderophore
	g004674	*ybtU*	Yersiniabactin siderophore
	g005682	*ybtT*	Yersiniabactin siderophore
	g004772	*ybtE*	Yersiniabactin siderophore
	g006209	*fyuA*	Yersiniabactin siderophore
LEE-encoded T3SS effectors/apparatus	g006831	*cesD2*	Chaperone
	g007976	*cesT*	Chaperone
	g003244	*cesF*	Chaperone
	g010160	*cesD*	Chaperone
	g001133	*espF*	LEE effector
	g006328	*espG*	LEE effector
	g004523	*espH*	LEE effector
	g001402	*espZ*	LEE effector
	g003196	*map*	LEE effector
	g005910	*etgA*	Peptidoglycan lytic enzyme
	g008063	*grlA*	Regulator
	g004082	*grlR*	Regulator
	g008050	*ler*	Regulator
	g003108	*sepL*	Secretion regulator
	g007990	*sepD*	Secretion regulator
	g018367	*escF*	T3SS apparatus
	g003794	*escD*	T3SS apparatus
	g004967	*sepQ*	T3SS apparatus
	g016340	*escP*	T3SS apparatus
	g008238	*escO*	T3SS apparatus
	g003334	*escN*	T3SS apparatus
	g000494_2	*escV*	T3SS apparatus
	g005225	*escI*	T3SS apparatus
	g007837	*escJ*	T3SS apparatus
	g007238	*escC*	T3SS apparatus
	g006399	*escU*	T3SS apparatus
	g005719	*escT*	T3SS apparatus
	g008348	*escS*	T3SS apparatus
	g006613	*escR*	T3SS apparatus
	g006614	*escL*	T3SS apparatus
	g006645	*escK*	T3SS apparatus
	g001845	*tir*	Translocated intimin receptor
	g004941	*espB*	Translocator

	g003099	*espD*	Translocator
	g003549	*espA*	Translocator
Non-LEE encoded T3SS effectors	g004187	*cif*	Cell-cycle-inhibitory factor
	g000997	*espFu*	Non-LEE-encoded effector protein
	g003189	*espJ*	T3SS effector
	g000915_1	*espK*	T3SS effector
	g003352	*espL2*	T3SS effector
	g001914_2	*espM1*	T3SS effector
	g001914_1	*espM2*	T3SS effector
	g004690	*espN*	T3SS effector
	g000370_2	*espR1*	T3SS effector
	g000683_1	*espR3*	T3SS effector
	g002657_1	*espV*	T3SS effector
	g005598	*espW*	T3SS effector
	g001351	*espX2*	T3SS effector
	g000929	*espX6*	T3SS effector
	g001816_1	*espX7*	T3SS effector
	g000466	*espY1*	T3SS effector
	g001738	*espY2*	T3SS effector
	g002724	*espY3*	T3SS effector
	g000684	*espY4*	T3SS effector
	g000913	*espY5*	T3SS effector
	g001104	*nleA*	T3SS effector
	g001411	*nleB*	T3SS effector
	g004348	*nleC*	T3SS effector
	g009605	*nleD*	T3SS effector
	g006587	*nleE*	T3SS effector
	g006668	*nleF*	T3SS effector
	g001825_2	*nleG-1*	T3SS effector
	g000526_3	*nleG-2*	T3SS effector
	g000526_1	*nleG2-1*	T3SS effector
	g000526_2	*nleG2-2*	T3SS effector
	g007752	*nleG5*	T3SS effector
	g001329_1	*nleG6*	T3SS effector
	g005755_1	*nleG7*	T3SS effector
	g001825_1	*nleG8-2*	T3SS effector
	g002513_1	*nleH1-1*	T3SS effector
	g002513_2	*nleH1-2*	T3SS effector
	g000398_2	*lifA/efa1*	Adherence factor
Regulation	g001414_1	*aggR*	Transcriptional activator of aggregative adherence fimbria
ABC transporter	g004308	*aatP*	ABC transporter for dispersin
	g003499	*aatB*	ABC transporter for dispersin
	g002288	*aatD*	ABC transporter for dispersin
T6SS apparatus	g000469_1	*aec29*	Chaperone
	g017162	*aaiM*	T6SS apparatus
	g002864	*aec24*	T6SS apparatus
	g000420_2	*aec25*	T6SS apparatus
	g001415	*aec26*	T6SS apparatus
	g000420_3	*aec32*	T6SS apparatus

	g000420_1	*aec31*	T6SS apparatus
	g001283	*aec23*	T6SS-associated gene
	g000469_2	*aec28*	T6SS-associated gene
Toxin	g001699_1	*hlyD*	Alpha-hemolysin
	g001349	*hlyB*	Alpha-hemolysin
	g000739_1	*hlyA*	Alpha-hemolysin
	g002368	*hlyC*	Alpha-hemolysin
	g007618	*cdtA*	Cytolethal distending toxin
	g006506	*cdtB*	Cytolethal distending toxin
	g007875	*cdtC*	Cytolethal distending toxin
	g007727_1	*eltA*	Heat-labile enterotoxin
	g005966_1	*eltB*	Heat-labile enterotoxin
	g004650	*estIa*	Heat-stable enterotoxin 1 (EAST1)

[a] Tag of the gene in the pangenome

[b] Annotated gene name by using the reference sequences of VFDB (Liu *et al.*, 2019)

[c] Annotated virulence factor by using the reference sequences of VFDB (Liu *et al.*, 2019). T3SS, type III secretion systems; LEE, locus of enterocyte effacement; ABC, ATP-binding cassette; T6SS, type VI secretion systems

II-3-2. 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. II-2A). 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. II-2A). 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. II-2B). 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. II-2B). The models using the GMM algorithm also performed poorly in discriminating the clinical and environmental isolates with a maximum accuracy of 44% (Fig. II-3). These results indicate that the unsupervised ML algorithms cannot effectively discriminate between the clinical and environmental isolates.

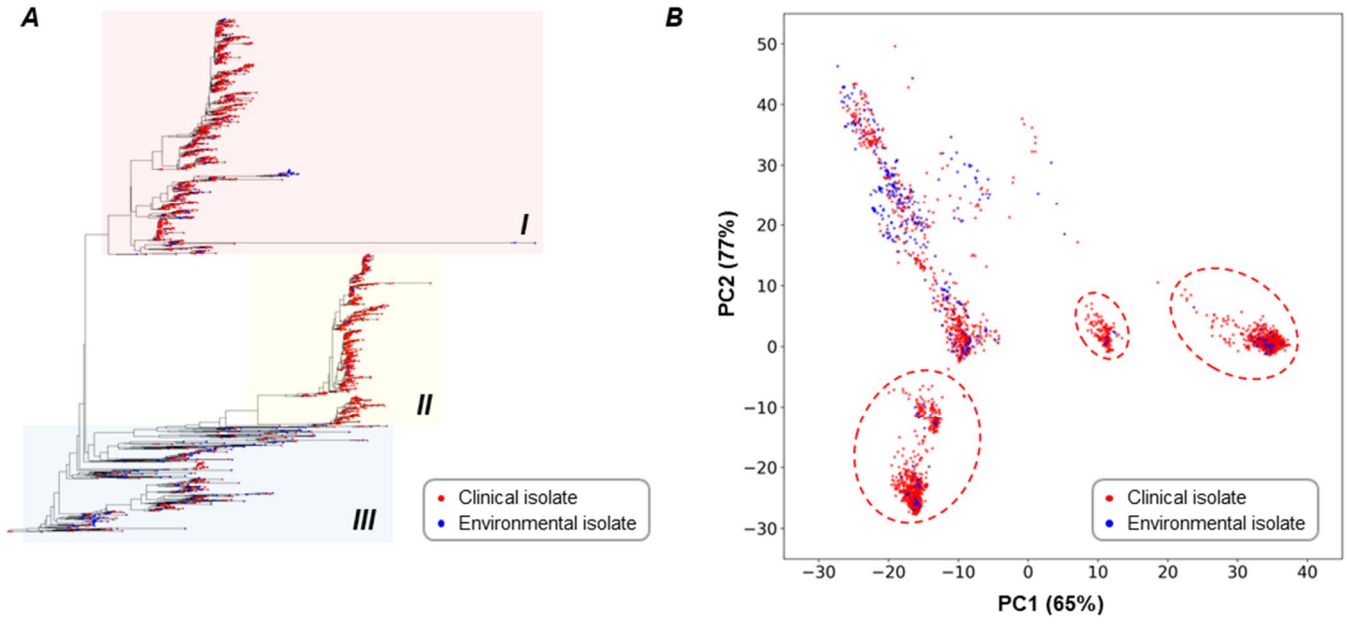

Figure II-2. Analyses of the EHEC isolates using the input dataset based on the unsupervised ML algorithms. The red dot represents the clinical isolate, and the blue dot represents the environmental isolate. (A) The phylogenetic tree of the EHEC isolates based on a maximum likelihood method. The three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

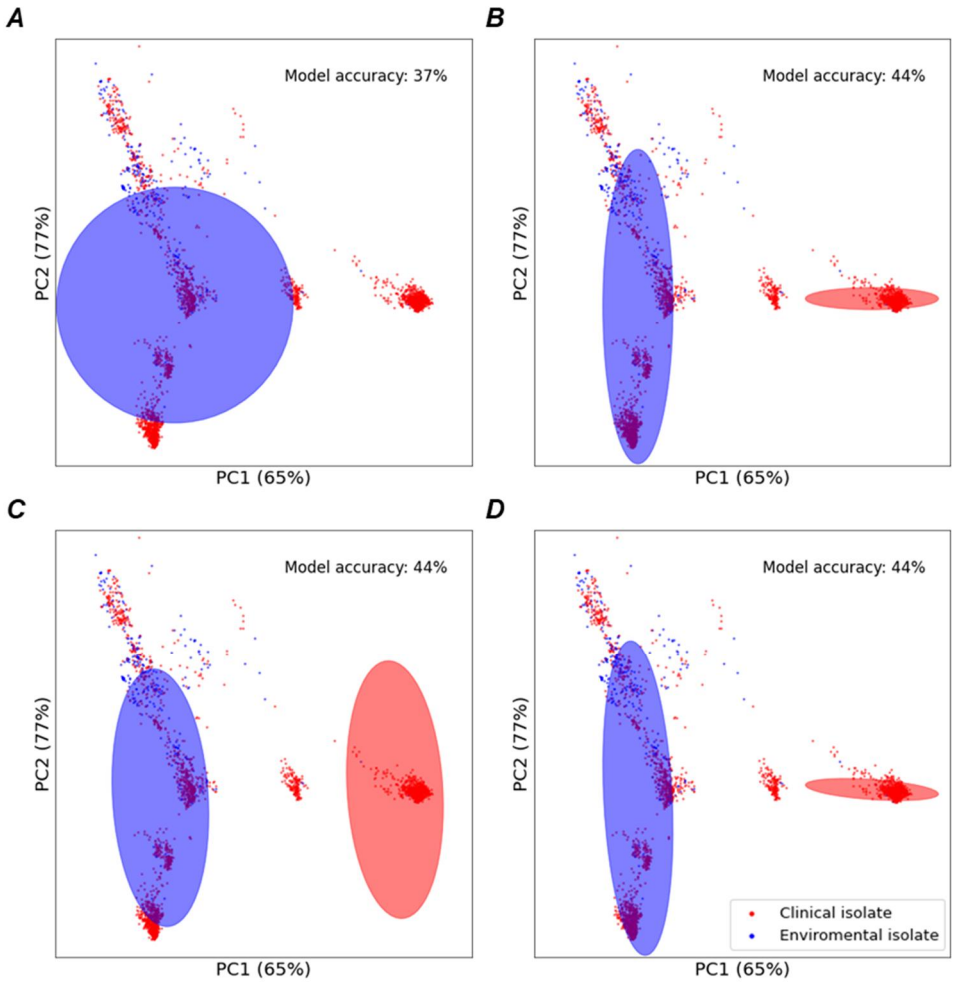

Figure II-3. The PCA plots of the clinical and environmental EHEC isolates. The clusters generated by GMM models for the clinical and environmental isolates are indicated by the red and blue circles, respectively. The covariance type of the GMM model for each plot is spherical (A), diagonal (B), tied (C), and full (D), respectively. The accuracy scores of the GMM models are noted at the top right of each plot. PC1, Principal component 1; PC2, Principal component 2.

II-3-3. 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 (Fig. II-4). These results indicated that the supervised ML models were able to discriminate between the clinical and environmental isolates. The MCC and 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. II-5A). The SVM model also presented the steepest receiver operating characteristic (ROC) curve with an area under the curve (AUC) score of 0.93 (Fig. II-5B and Fig. II-6), 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 (Fig. II-7 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 EHEC isolates.

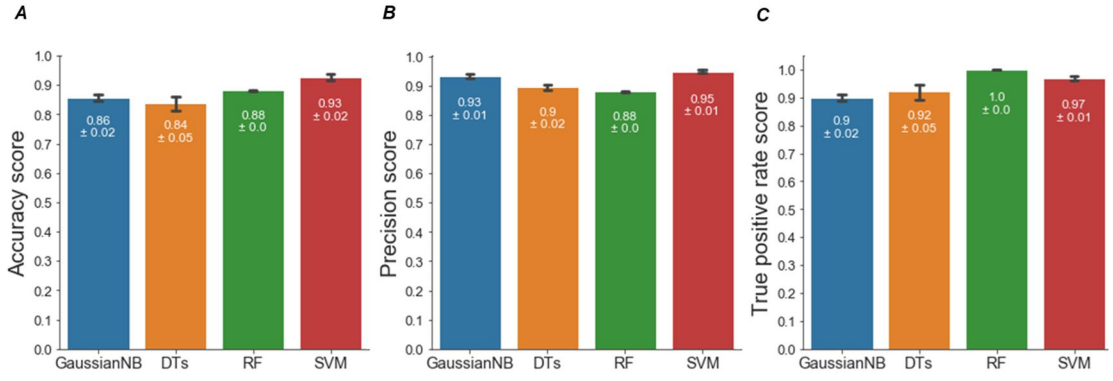

Figure II-4. The discrimination performances of the supervised ML models for the EHEC isolates in the input. The supervised ML models using four different algorithms: GaussianNB, DTs, RF, and SVM, as indicated. The accuracy (A), precision (B), and true positive rate (C) scores of the indicated supervised ML models are presented as bar plots. The average scores of the individual models are indicated at the tip of the bars. SD is represented by the error bar and score.

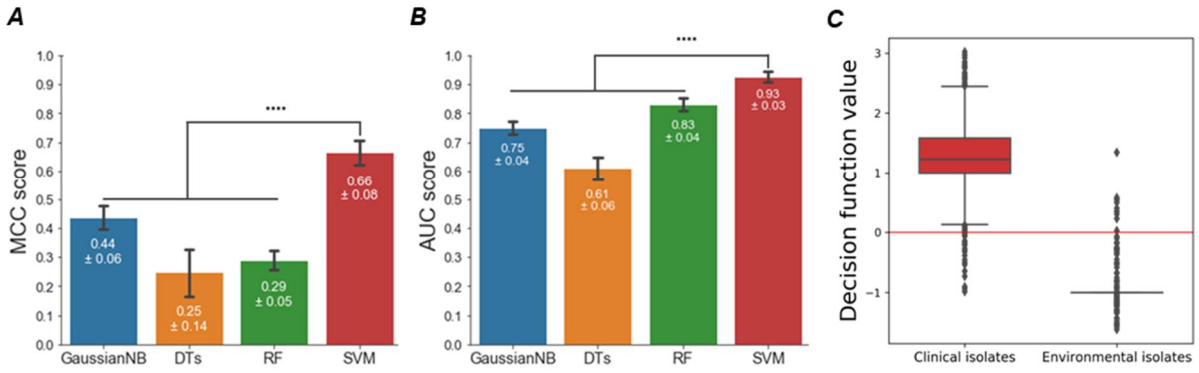

Figure II-5. The discrimination performances of the supervised ML models for the EHEC isolates in the input dataset: MCC, AUC, and decision function values. (A and B) The bar plots of the discrimination performances of the supervised ML models using four different algorithms: GaussianNB, DTs, RF, and 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.

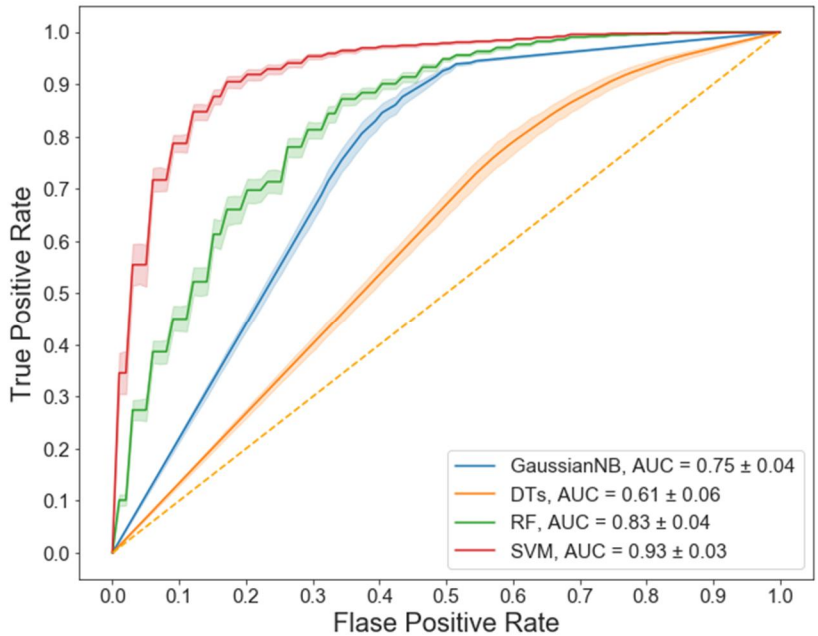

Figure II-6. The ROC curves of the supervised ML models using four different algorithms. The average AUC scores of each model are presented in the legend. Errors of the curves and AUC scores represent the SD.

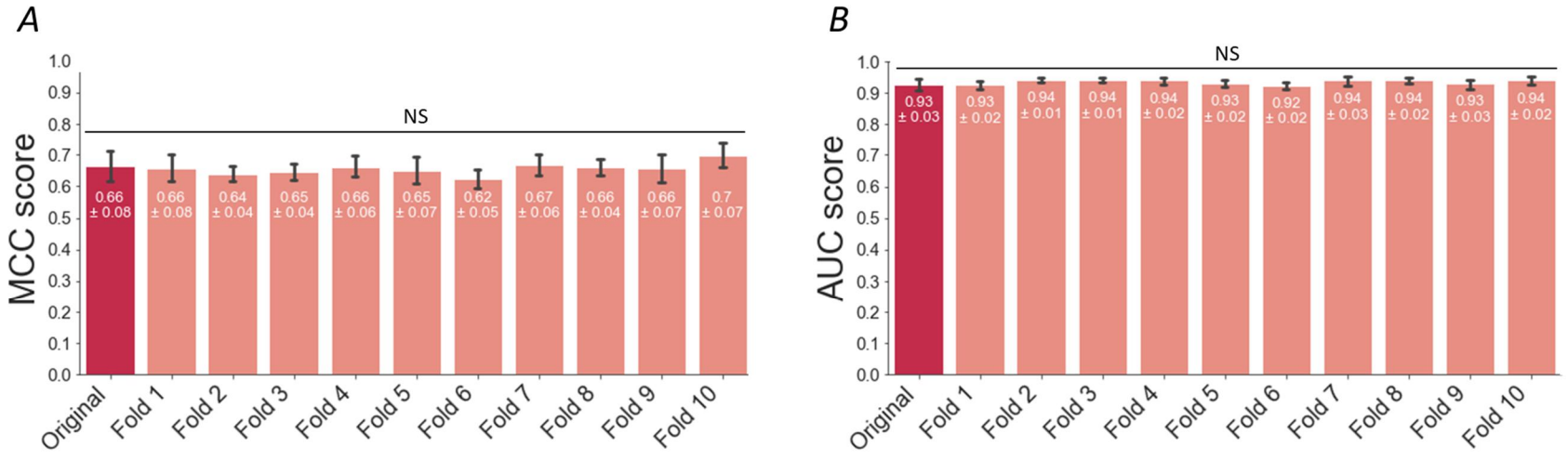

Figure II-7. The discrimination performances of the SVM models scored with MCC (A) and AUC (B). The scores of the models are presented as bar plots. SD is represented by the error bar and score. The average scores of the individual models are indicated at the tip of the bars. Statistical significance of the scores was determined by one way-ANOVA (NS, not significant). The datasets used to train the models were labeled under the graph. Original, the input dataset; Fold, the dataset consisting of the significant genes from the training set of each fold of the stratified 10-fold CV.

II-3-4. The SVM model evaluates the pathogenic potential of the EHEC 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. II-5C, 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 EHEC isolates using the input dataset.

II-3-5. The SVM model evaluates the pathogenic potential of the EHEC isolates according to their source attribution and clinical outcomes

The environmental isolates from cattle, dairy products, and farm products, the major sources of EHEC 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. II-8A). 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. II-5C). 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. II-8A). The SVM model effectively estimated that many of the environmental isolates from the major sources of the EHEC outbreaks are pathogenic as previously reported by the source attribution of EHEC (World Health Organization, 2011, 2018; Koutsoumanis *et al.*, 2020), 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 EHEC 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. II-8B). This result indicates that the SVM model correctly evaluates the pathogenic potential of the EHEC isolates consistent with their clinical outcomes. Altogether, the combined results suggest that the SVM model is able to produce an effective and reliable assessment of the pathogenic potential of EHEC isolates using only their significant gene profiles extracted from the WGS data.

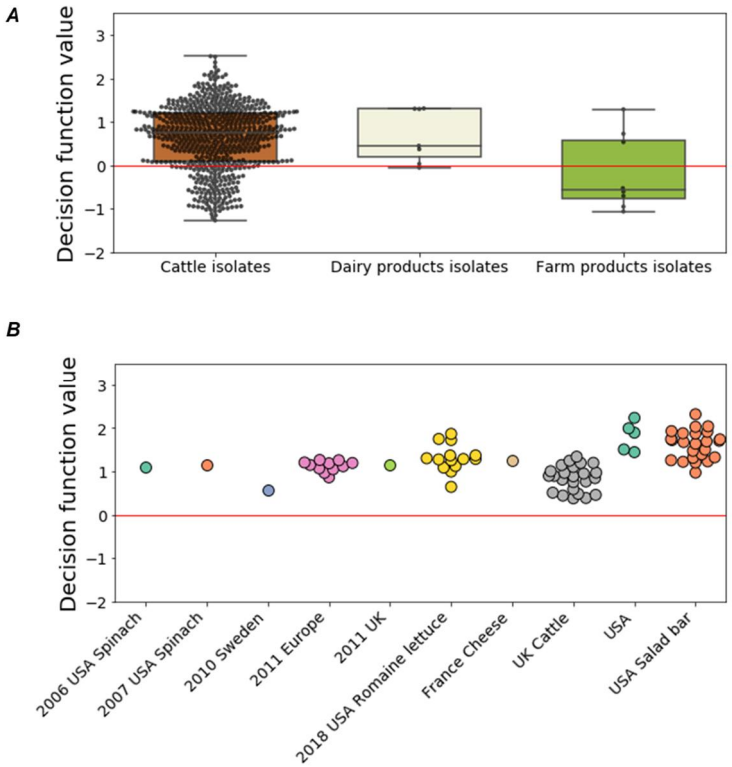

Figure II-8. The box and swarm plots of the decision function values of the isolates associated with the EHEC outbreaks. (A) The box and swarm plots of the decision function values of the isolates from cattle, dairy products, and farm products that were excluded from the negative control group. Each 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.

II-3-6. 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 EHEC isolates with O26, O157, O121, O145, O111, O104, O91, O103, and O55 serotypes are pathogenic (European Food Safety Authority, 2013; Gould *et al.*, 2013; Eichhorn *et al.*, 2015; World Health Organization, 2018; Koutsoumanis *et al.*, 2020). 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. II-9A and Table II-6), comparable with the clinical isolates (Fig. II-5C). In contrast, the decision function values of Q1 for the isolates with the O91, O103, and O55 serotypes were between 1 and 0 (Fig. II-9A and Table II-6), 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 EHEC 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. II-9A and Table II-6). 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. II-9B and Table II-6), 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. II-9B and Table II-6), 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. II-9B and Table II-6), indicating that most of these isolates may have high pathogenic potential. Accordingly, the SVM model successfully predicted the pathogenic potential of the EHEC 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. The virulence gene combination method showed that the EHEC isolates carrying a combination of *stx2a* + *eae* or *aggR* are highly pathogenic (World Health Organization, 2018; Koutsoumanis *et al.*, 2020). The SVM model revealed that the isolates with *stx2a* + *eae* or *aggR* had decision function values of Q1 over 1.00 and were pathogenic (Fig. II-9C). The isolates only with *stx2a* or *stx2d*, however, showed a broad spectrum of decision function values ranging from −1.61 to 2.38 (Fig. II-9C), 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 EHEC isolates. Moreover, about 56% of the isolates in the input dataset (1,504/2,646) do not have such gene combination as *stx2a* + *eae* or *aggR*, *stx2a*,or *stx2d*, 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 EHEC isolates.

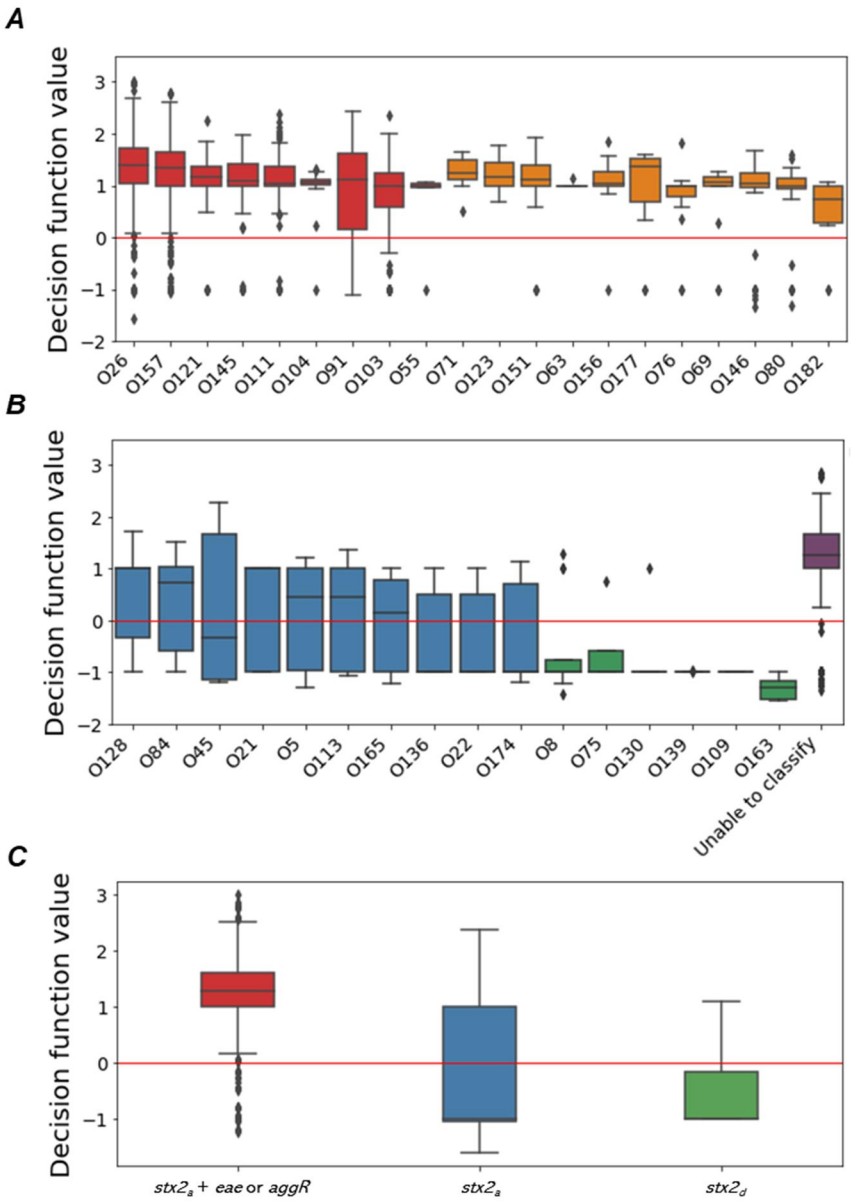

Figure II-9. The box plots of the decision function values of the EHEC isolates in the input dataset grouped by the conventional assessment methods. (A and B) The box plots of the decision function values of the isolates grouped by serotypes. The serotype groups with decision function values of Q1 over 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 Table II-6. (C) The box plots of the decision function values of the isolates grouped by virulence gene combinations. The *stx2a* + *eae* or *aggR* group, *stx2a* group, and *stx2d* 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.

Table II-6. Median, Q1, and Q3 values of decision function values of EHEC isolates grouped by serotypes.

Serotype	Median	Q1	Q3
O26	1.40	1.05	1.72
O157	1.33	1.00	1.65
O121	1.16	1.00	1.37
O145	1.09	1.00	1.42
O111	1.04	1.00	1.36
O104	1.06	1.01	1.12
O91	1.11	0.15	1.63
O103	1.00	0.57	1.24
O55	1.00	0.97	1.03
O71	1.24	1.11	1.49
O123	1.16	1.00	1.44
O151	1.12	1.00	1.38
O63	1.00	1.00	1.00
O156	1.04	1.00	1.26
O177	1.35	0.69	1.52
O76	1.00	0.79	1.00
O69	1.06	1.00	1.15
O146	1.03	0.96	1.25
O80	1.00	0.94	1.14
O182	0.73	0.28	1.00
O128	1.00	-0.34	1.01
O84	0.74	-0.6	1.03
O45	-0.32	-1.14	1.67
O21	1.00	-1.00	1.00
O5	0.44	-0.95	1.00
O113	0.45	-1.00	1.00
O165	0.16	-1.00	0.79
O136	-1.00	-1.00	0.5
O22	-1.00	-1.00	0.5
O174	-1.00	-1.00	0.71
O8	-1.00	-1.00	-0.75
O75	-1.00	-1.00	-0.58
O130	-1.00	-1.00	-1.00
O139	-1.00	-1.00	-1.00
O109	-1.00	-1.00	-1.00
O163	-1.3	-1.51	-1.17
Unable to classify	1.27	1.00	1.68

Q1, first quartile; Q3, third quartile

II-3-7. Permutation importance analyses identify the genes important to estimate the pathogenicity of the EHEC 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. II-10A). 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. II-10B). Only four of the top 20 important genes carry the previously reported functions: antitermination protein Q gene (Brüssow *et al.*, 2004; Steyert *et al.*, 2012), antitermination protein Q homolog gene *quuQ* (Brüssow *et al.*, 2004; Steyert *et al.*, 2012), non LEE-encoded effector protein gene *espFu* (Martins *et al.*, 2017, 2020), and toxin-antitoxin (TA) system gene *lsoA* (Otsuka and Yonesaki, 2012) (Fig. II-10A and Table II-7). 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. II-10 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 (Zatyka and Thomas, 1998), sialic acid catabolism *nan* genes (Kalivoda *et al.*, 2013), TA system genes *phd-doc* (Lehnherr *et al.*, 1993), and osmotic stress response gene *mscS* (Bremer and Krämer, 2019) (Fig. II-10C and

Table II-8). Consequently, the permutation importance analyses of the input dataset identified the genes important to estimate the pathogenicity of the EHEC 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 EHEC isolates.

Meanwhile, *in silico* analysis showed that about 12% of input dataset genes (406/3,462) were predicted as phage genes. However, the ratio of phage genes in the genes of top 25% importance from the individual gene level permutation importance analysis is about 25% (198/866) (Figure II-11A). The ration of phage genes in the important genes from the clustered gene level permutation importance analysis is about 23% (231/950) (Figure II-11B). These results indicated that the proportion of phage genes observed in the important genes is higher than that in the whole input dataset genes, supporting the claim that the phages contribute to pathogenicity of EHEC (Koutsoumanis et al., 2020).

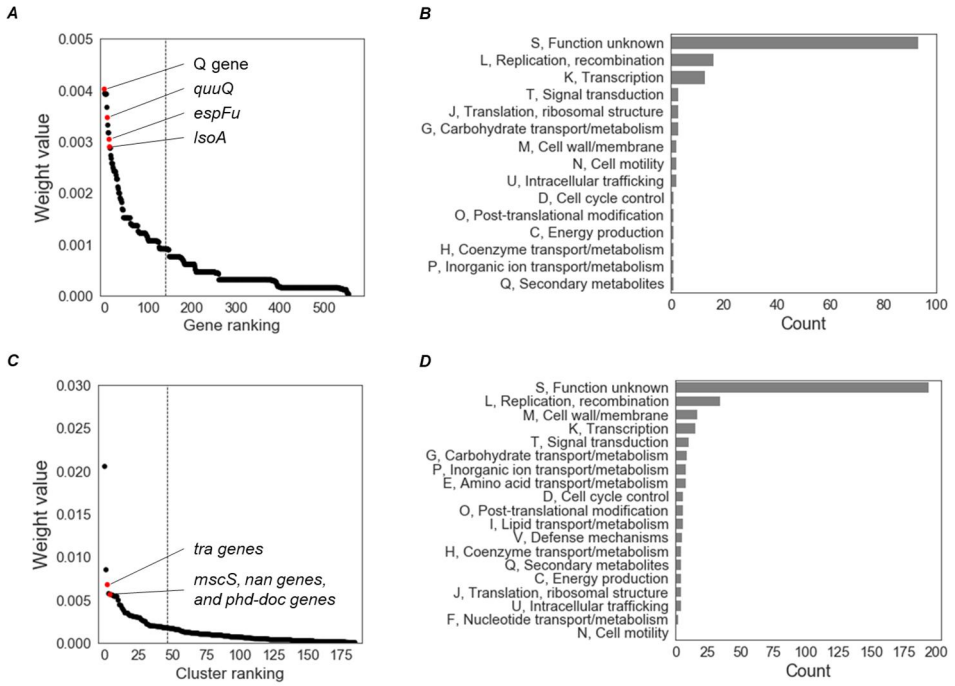

Figure II-10. Permutation importance analyses of the input dataset. (A and C) The scatter plots of the importance of individual genes (A) and correlated gene clusters (C). The importance of each gene and cluster is presented by the weight value. The important genes and clusters are plotted by the rank 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.

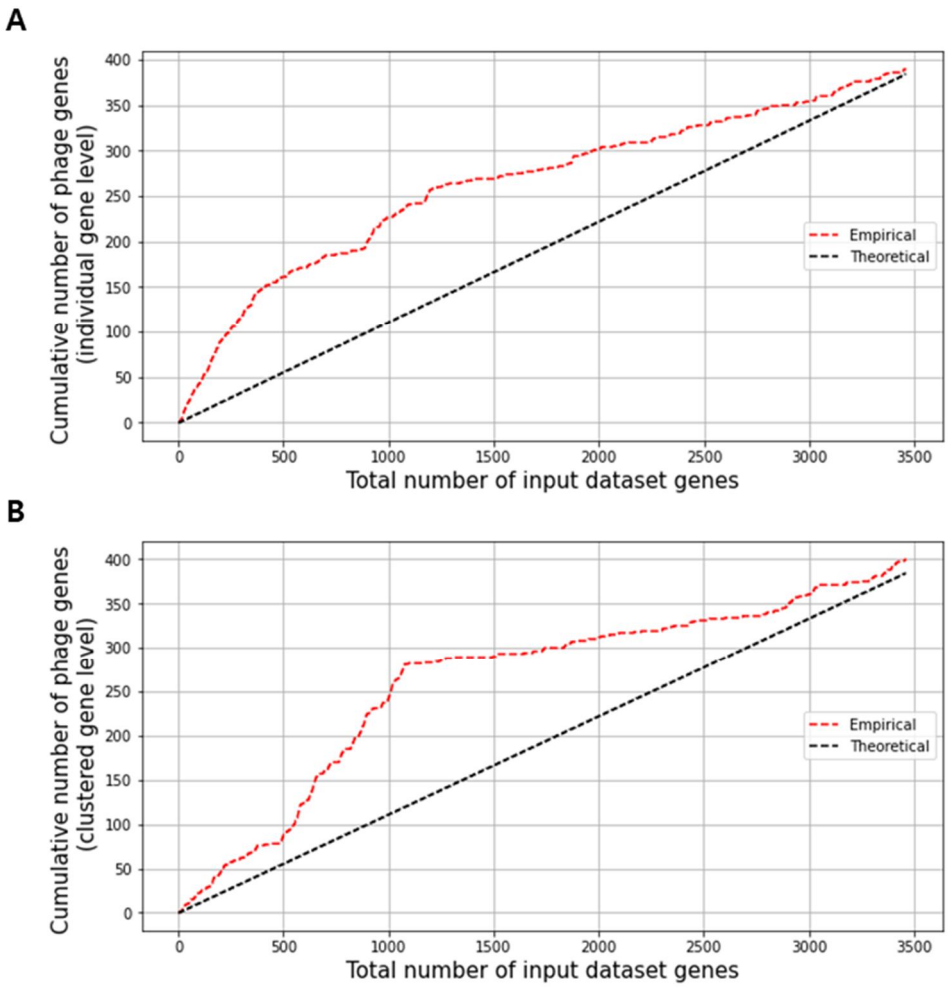

Figure II-11. The cumulative number of phage genes according to importance rank of EHEC genes resulted from the individual- and clustered-gene level permutation importance analysis. The dotted red line indicates the cumulative number of phage genes observed empirically. The dotted black line indicates the cumulative number of phage genes calculated theoretically based on the average ratio of phage genes in the input dataset genes.

Table II-7. Top 20 important genes in the permutation importance analysis of the input dataset.

Gene tag[a]	Weight[b]	SD[c]	Gene name[d]	Protein name[e]
g002124_1	0.0040	0.0019		Antitermination protein Q
g007884	0.0039	0.0032		hypothetical protein
g003829	0.0039	0.0007		hypothetical protein
g005099	0.0039	0.0007		hypothetical protein
g004264	0.0039	0.0007		hypothetical protein
g004621	0.0039	0.0007		hypothetical protein
g006123	0.0037	0.0022		hypothetical protein
g000997	0.0035	0.0022	*espFu*	Non-LEE-encoded effector protein
g006049	0.0033	0.0018		hypothetical protein
g011033	0.0032	0.0016		hypothetical protein
g001239_1	0.0032	0.0020		hypothetical protein
g013010	0.0030	0.0007	*lsoA*	mRNA endoribonuclease LsoA
g003934	0.0029	0.0017	*quuQ*	Prophage antitermination protein Q homolog
g008365	0.0029	0.0005		hypothetical protein
g002431_1	0.0029	0.0005	*yaiX*	Putative uncharacterized acetyltransferase
g006097_1	0.0027	0.0014		hypothetical protein
g008512_1	0.0027	0.0020		hypothetical protein
g000970_1	0.0026	0.0010	*ydeR*	Uncharacterized fimbrial-like protein
g000812_2	0.0026	0.0007		hypothetical protein
g004651	0.0026	0.0007		hypothetical protein

[a] Tag of the gene in the pangenome data

[b] Weight value of gene in the permutation importance analysis

[c] SD, standard deviation

[d] Annotated gene name by using the reference sequences of UniProtKB (UniProt Consortium and Bateman, 2019)

[e] Annotated protein name by using the reference sequences of UniProtKB (UniProt Consortium and Bateman, 2019)

Table II-8. The genes included in the top 5 important clusters in the permutation importance analysis of the input dataset.

Cluster No.	Gene tag[a]	Weight[b]	SD[c]	Gene name[d]	Protein name[e]
443	g009532	0.0205	0.0016		WecB/TagA/CpsF family glycosyltransferase
	g013314				Glucosyltransferase family
	g011463				hypothetical protein
	g012310				hypothetical protein
	g012328				hypothetical protein
	g013916				hypothetical protein
	g006064				hypothetical protein
	g007217				hypothetical protein
	g008660				hypothetical protein
	g018987				hypothetical protein
	g019644				hypothetical protein
	g018892				hypothetical protein
	g009713				hypothetical protein
	g010650				hypothetical protein
77	g011033	0.0085	0.0018		hypothetical protein
	g000901_2				hypothetical protein
	g005792				hypothetical protein
	g000764_2				hypothetical protein
415	g006176	0.0067	0.0014	*traG*	Conjugal transfer protein TraG
	g000200_2			*traD*	Conjugal transfer protein TraD
	g001196_2				hypothetical protein
360	g010141	0.0057	0.0019		Peptidoglycan hydrolase gp27
	g009929				Uncharacterized HNH endonuclease L245
	g001140				Putative type VI secretion system protein
	g005946				hypothetical protein
	g003262				hypothetical protein
	g003400				hypothetical protein
	g007070				hypothetical protein
190	g003906_2	0.0056	0.0033	*nanA*	N-acetylneuraminate lyase
	g005067_2			*nanC*	Probable N-acetylneuraminic acid outer membrane channel protein NanC
	g000887_3			*nanK2*	N-acetylmannosamine kinase

	g001278_2			*nanT2*	Sialic acid transporter NanT 2
	g003496_2			*nanR*	HTH-type transcriptional repressor NanR
	g003593_2			*doc*	Toxin Doc
	g010862			*phd*	Antitoxin Phd
	g002121_2			*mscS*	Small-conductance mechanosensitive channel
	g002115				Transposase for insertion sequence element IS1111A
	g003140				Alpha/beta hydrolase
	g000364				YadA family protein
	g012713				Probable transport protein
	g000693				Probable family 20 transposase
	g012628				Uncharacterized protein HI_0093
	g012152				Uncharacterized transporter HI_0092
	g018135				hypothetical protein
	g019448				hypothetical protein
	g000656_3				hypothetical protein
	g000939_2				hypothetical protein
	g011791				hypothetical protein
	g010598				hypothetical protein
	g010657				hypothetical protein
	g004606_1				hypothetical protein
	g004595_1				hypothetical protein
	g002204_2				hypothetical protein
	g002213_2				hypothetical protein
	g013944				hypothetical protein

[a] Tag of the gene in the pangenome

[b] Weight value of the cluster in the permutation importance analysis

[c] SD, standard deviation

[d] Annotated gene name by using the reference sequences of UniProtKB (UniProt Consortium and Bateman, 2019)

[e] Annotated protein name by using the reference sequences of UniProtKB (UniProt Consortium and Bateman, 2019)

II-4. Discussion

To develop the most proper ML model in evaluating the pathogenic potential of the EHEC isolates, various ML models were compared by their performances on discriminating between clinical and environmental isolates. In contrast to the unsupervised ML models (Fig. II-2 A and B, Fig. II-3), the supervised ML models successfully discriminated between the clinical and environmental isolates using the input dataset (Fig. II-4). 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. II-5 A and B, Fig. II-6). 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. II-5C). 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. II-5C and Fig. II-12A). However, the MLP model converged its sigmoid function values to 0 for the non-pathogenic isolates or 1 for the pathogenic isolates (Fig. II-12 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 (Fig. II-5C, Fig. II-8 A and B). Therefore, the SVM model was more appropriate to

estimate the pathogenic potential of EHEC 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 EHEC isolates, assuming that the isolates originating from humans or cattle are pathogenic or nonpathogenic, respectively (Lupolova *et al.*, 2016, 2017). 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 EHEC outbreaks. Instead, the positive and negative control groups were set 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 (Fig. II-1). In contrast to the previous ML model that estimated only under 10% of the cattle isolates to be pathogenic (Lupolova *et al.*, 2016, 2017). Our SVM model estimated about 80% of the cattle isolates to be pathogenic (Fig. II-8A). In addition, the SVM model also evaluated that many of the isolates from dairy and farm products are pathogenic (Fig. II-8A). These results also supported that cattle, dairy products, and farm products could be sources of the pathogenic EHEC (World Health Organization, 2011, 2018; Koutsoumanis *et al.*, 2020) and thereby should be handled with special care. Moreover, the SVM model correctly predicted the EHEC isolates with the history of outbreaks to carry high pathogenic potential (Fig. II-8B), indicating that the SVM

model prediction is 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 EHEC isolates.

The SVM model correctly classified the EHEC isolates previously designated as pathogenic by the conventional methods into the pathogenic group (Fig. II-9 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. II-9 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 EHEC 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. II-9B), revealing its broad applicability. Notably, many of these isolates are predicted to have high pathogenic potential (Fig. II-9B), emphasizing the necessity of the SVM model rather than conventional methods. The MLP model also correctly classified the EHEC isolates previously designated as pathogenic by the conventional methods into the pathogenic group (Fig. II-13 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 (Fig. II-13 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 EHEC 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. II-10 A and C, Table II-7 and II-8). The antitermination protein Q gene and its homolog gene *quuQ* participate in the regulation of Shiga toxin genes (Brüssow *et al.*, 2004; Steyert *et al.*, 2012). The non-LEE-encoded effector protein gene *espFu* is involved in the formation of attaching and effacing lesion, the major mechanism of EHEC infection (Martins *et al.*, 2017, 2020). 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 (Lehnherr *et al.*, 1993; Otsuka and Yonesaki, 2012). The *tra* genes are conjugal transfer system genes (Zatyka and Thomas, 1998). Considering that the horizontal transfer of plasmids is a major route of EHEC to acquire virulence factors (Caprioli *et al.*, 2005; Ogura *et al.*, 2007), these plasmid-encoded and plasmid transfer-related genes possibly contribute to the pathogenicity of EHEC. The *nan* genes are the sialic acid catabolism genes (Kalivoda *et al.*, 2013) and enable the pathogen to utilize the host sialic acids as nutrient sources (Kalivoda *et al.*, 2003, 2013), contributing to the survival and

pathogenesis in the host (Vimr, 2013). The *mscS* is an osmotic stress response gene (Bremer and Krämer, 2019) and is up-regulated when EHEC is exposed to the host intestinal environment (Pieper *et al.*, 2013). Altogether, these results indicate that the SVM model employs the genes associated with the pathogenesis of EHEC to estimate its pathogenic potential. However, most of the important genes have yet unknown functions (Fig. II-10 B and D). Nevertheless, new genes significantly associated with the pathogenicity of EHEC could be discovered from these important genes with unknown functions, further elucidating the pathogenicity of EHEC.

In conclusion, I developed an ML model using the SVM algorithm to effectively estimate the pathogenic potential of EHEC isolates using their significant gene profiles extracted from the WGS data, rendering it more extensively applicable than the conventional assessment methods. However, the ML-based approach has several limitations. First of all, the assumption that the environmental isolates are non-pathogenic is not fully supported by the experimental data, thus the pathogenic potential of the environmental isolates is needed to be further validated empirically. Second, the ML model has been learned to only predict the presence or absence of pathogenic potential, not to predict the severity of the pathogenicity. Third, the input dataset used to train the ML model only includes the genetic differences in the accessory genome of EHEC, not including those in the core genome of EHEC caused by single nucleotide polymorphisms (SNPs) or indel. Finally, since the phenotypic variation is associated with complex interactions of gene transcription and

protein translation, there is a fundamental limit to accurately predicting the pathogenic phenotype of EHEC isolates using their partial genetic information. Nevertheless, this study presents a novel approach to predict the pathogenic potential of EHEC isolates using genome data based on the source attribution and ML algorithms. This ML-based approach could be applied to other pathogens and be used to identify the potential risk of newly emerging pathogens.

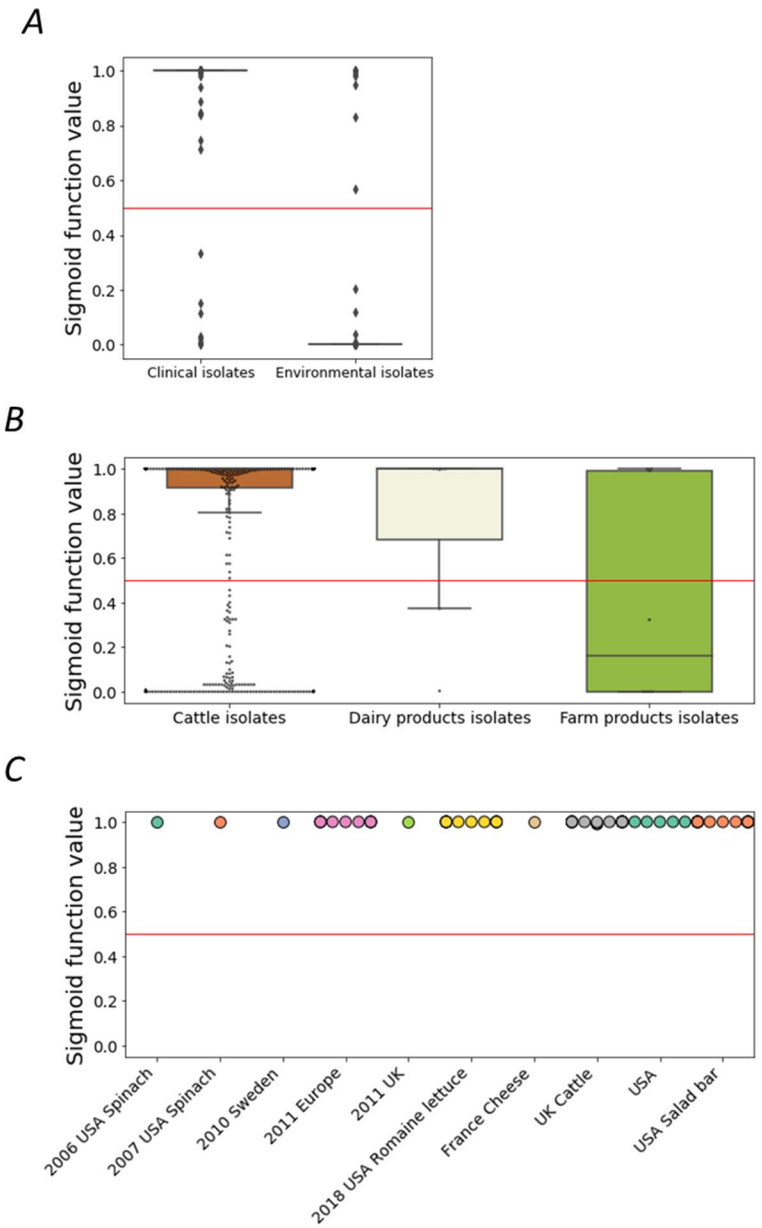

Figure II-12. The evaluation performances of the MLP model. The sigmoid function values of the isolates were calculated by the MLP model. The isolates with a sigmoid function value over 0.5 (above the red line) or under 0.5 (below the red line) are classified into the pathogenic or nonpathogenic group, respectively. The end

lines of each box show the Q1 and Q3 of the values. (A) The box plots of the sigmoid function values of the clinical and environmental isolates in the input dataset. The clinical and environmental isolates had sigmoid function values of median 1.00 (Q1, Q3: 1.00, 1.00) and 0.00 (Q1, Q3: 0.00, 0.00), respectively. (B) The box and swarm plots of the sigmoid function values of the isolates from cattle, dairy products, and farm products. (C) The swarm plots of the sigmoid 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.

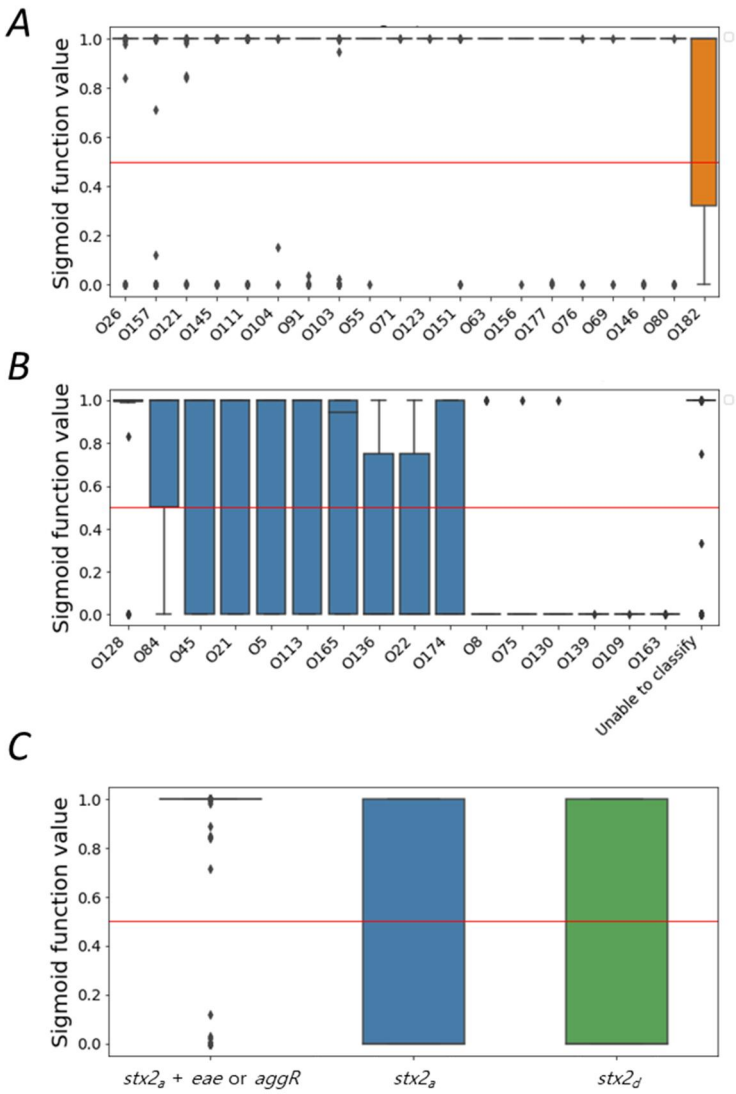

Figure II-13. The box plots of the sigmoid function values of the EHEC isolates in the input dataset grouped by the serotype and the virulence gene combination method. The isolates with a sigmoid function value over 0.5 (above the red line) or under 0.5 (below the red line) are classified into the pathogenic or nonpathogenic group, respectively. The end lines of each box show the Q1 and Q3 of the values. (A and B) The box plots of the sigmoid function values of the isolates grouped by

serotypes. The group of isolates that cannot be classified according to their serotype is labeled as ‘Unable to classify.’ (C) The box plots of the sigmoid function values of the isolates grouped by virulence gene combinations.

Chapter III.

Independent component analysis identifies the modulons expanding the transcriptional regulatory networks of enterohemorrhagic *Escherichia coli*

Part of this work in Chapter III was published in *Frontiers in Microbiology* in 2022, as an article entitled "Independent component analysis identifies the modulons expanding the transcriptional regulatory networks of enterohemorrhagic *E. coli*".

III-1. Introduction

Transcriptional regulatory networks (TRNs) regulate the expression of the target genes for the pathogens to adapt to various environments. The understanding of TRNs and their target genes enables the prediction of molecular mechanisms by which pathogens cause disease and survive under host-specific conditions (Karmali, 2017). Advances in next-generation sequencing technologies facilitate analyzing the large-scale RNA-Seq and comparing the transcriptome of the pathogens grown under specific conditions or lacking a particular transcription factor(s) (TF) (Westermann *et al.*, 2012; DuPai *et al.*, 2020). However, the transcriptome data obtained from the genes expressed under specific experimental conditions or by a certain TF are still limited to comprehensively understand the TRNs and their target genes (Sastry *et al.*, 2019; DuPai *et al.*, 2020). Therefore, to overcome this limitation, studies have been performed to analyze bioinformatically the large-scale transcriptome data of the pathogens and to define the modulons, the independent sets of genes co-regulated under various conditions regardless of their genetic backgrounds (Saelens *et al.*, 2018; Sastry *et al.*, 2019; DuPai *et al.*, 2020; Tan *et al.*, 2020).

Enterohemorrhagic *Escherichia coli* (EHEC) causes a broad spectrum of human illnesses ranging from mild diarrhea to hemolytic uremic syndrome, often leaving permanent damage to the kidney (Karmali, 2017). The TRNs of Ler and Shiga toxin

(Stx) prophage encoding the major virulence factors of EHEC have been studied extensively to understand the molecular pathogenesis of the pathogen. Ler, encoded by *ler*, regulates the locus of enterocyte effacement (LEE) genes necessary to form attaching and effacing (AE) lesions, the central pathogenesis of EHEC (Kenny *et al.*, 1997; Mellies *et al.*, 1999; Elliott *et al.*, 2000; Tobe *et al.*, 2006). Ler also regulates the genes encoding non-LEE-encoded effector (Nle) proteins crucial for forming A/E lesions (Kelly *et al.*, 2006; Li *et al.*, 2006; Tobe *et al.*, 2006), demonstrating that the Ler TRN contains additional non-LEE genes. Additionally, the Stx prophage TRNs include *stx1* and *stx2* of EHEC, located in the CP-933V and BP-933W prophages, respectively (Xu *et al.*, 2012). The expressions of the Stx genes are regulated by the antiterminator Qs which allows the transcription of the prophage genes by preventing the formation of intrinsic terminators in their promoters (Casjens and Hendrix, 2015; Sy *et al.*, 2020).

Meanwhile, the TRNs also coordinate the expressions of the target genes for pathogens to survive under various growth conditions by recognizing the changes in the environmental signals. For example, the copper transport TRNs of EHEC consisting of *cusCFBA* involved in the detoxification of toxic heavy metals are induced by the high copper ions (Delmar *et al.*, 2015). The *cusCFBA* are generally suppressed by the global regulator H-NS encoded by *hns* in the absence of the heavy metal ions (Atlung and Ingmer, 1997; Lang *et al.*, 2007). Conversely, the target genes of certain TRNs also could be suppressed by the environmental signals. For example,

The LEE genes of the Ler TRN are suppressed in the presence of indole, synthesized by the tryptophanase encoded by the *tna* (Kumar and Sperandio, 2019). Similarly, the TRNs containing *thiBP* and *thiCEFGH* involved in the thiamine transport and biosynthesis, respectively, are also suppressed in the presence of thiamine (Vander Horn *et al.*, 1993; Webb *et al.*, 1998; Miranda-Rios *et al.*, 2001).

In this study, independent component analysis (ICA), a machine learning method that decomposes a mixture of components into the independent components (James and Hesse, 2005; Sastry *et al.*, 2019; Tan *et al.*, 2020), was used to decompose the large-scale transcriptome data of EHEC into the sets of independent modulon, which contains the target genes of several TRNs. The LEE and the Stx modulons mainly consisted of the target genes of the Ler and the Stx prophage TRNs, respectively, indicating that ICA properly grouped the sets of the co-regulated genes of EHEC into the modulons. Further investigation identified the Z0395 gene and the *thi* and *cus* locus genes as novel element genes of the LEE and Stx modulons, respectively. Accordingly, the Stx prophage genes were also regulated by thiamine and copper ions known to control the *thi* and *cus* locus genes, respectively. Changed expressions of the modulons consisting of the inherently co-regulated genes also successfully explained the differential expressions of the virulence and survival genes of EHEC during the course of infection in bovine.

III-2. Materials and Methods

III-2-1. Generation of the trimmed transcriptome data of EHEC

The raw-sequencing reads of available RNA-Seq data of EHEC were retrieved from the Sequence Read Archive (SRA) database at the National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/sra) (Table III-1). The reads were mapped to the reference genome of EHEC EDL933 (AE005174.2) using Spliced Transcripts Alignment to a Reference (STAR) (Dobin *et al.*, 2013). The reads aligned to the reference genome were counted using the HTSeq (Anders *et al.*, 2015). The genes with under ten fragments per million-mapped reads across the whole RNA-Seq data were removed before further analyses to ensure the quality of the data. The raw read counts were normalized using the trimmed mean of *M* values (TMM) method from the R *edgeR* package (Robinson and Oshlack, 2010; Robinson *et al.*, 2010). The normalized data with $R^2 < 0.9$ between biological replicates were discarded to trim the technical noise (Fig. III-1A). The trimmed transcriptome data were log-transformed ($\log_2$ TMM+1) for further analysis (Table III-2).

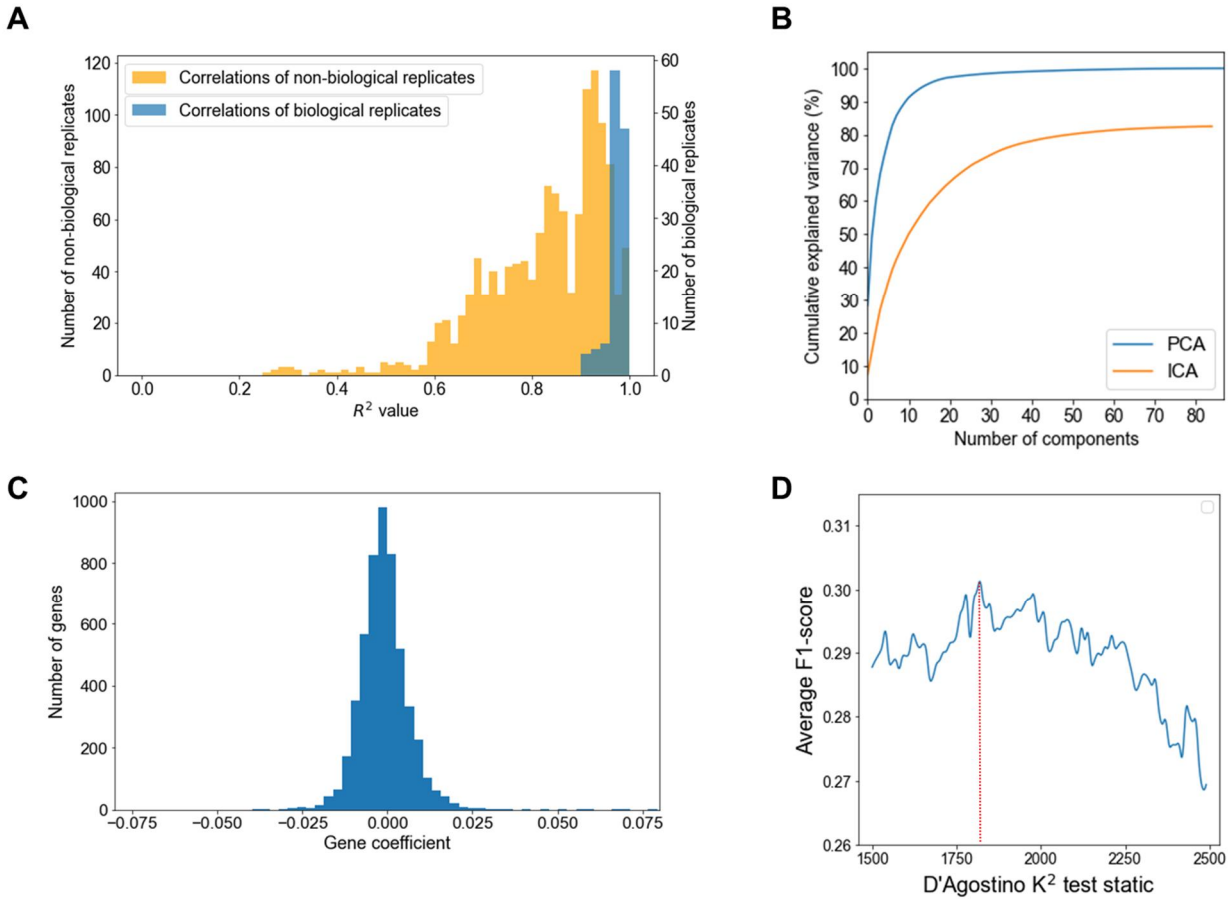

Figure III-1. Summary of the data processing. (A) Histogram of the correlation values between the biological replicates or non-biological

replicates of the transcriptome data of EHEC. The correlations between the replicates were calculated using R^2 values. Blue bars represent the correlations between the biological replicates, and the orange bars represent the correlations between non-biological replicates. (B) Cumulative explained variance (CEV) for the 88 transcriptome data of EHEC calculated by using the principal component analysis (PCA) (blue) and the independent component analysis (ICA) (orange). The independent gene components identified by using the ICA explained 83% of the total expression variance of the 88 transcriptome data. (C) Histogram of the gene coefficients in an independent gene component. The gene coefficients in any independent gene component display the histogram with a similar distribution. Most of the gene coefficients are near zero. (D) Average F1 scores calculated under the varied D'Agostino K^2 statistic cutoff ranging from 1,500 to 2,500 with an increment of 10. Red dotted line indicates the optimal cutoff value, 1800.

Table III-1. Detailed experimental conditions of the transcriptome data

SRR number or condition	Strain	Media	Detailed condition
ERR1370908	EDL933	Ground beef	EDL933 in ground beef with microflora
ERR1370909	EDL933	Ground beef	EDL933 in ground beef with microflora
ERR1370910	EDL933	Ground beef	EDL933 in ground beef with microflora
ERR1370911	EDL933	Ground beef	EDL933 in ground beef without microflora
ERR1370912	EDL933	Ground beef	EDL933 in ground beef without microflora
ERR1370913	EDL933	Ground beef	EDL933 in ground beef without microflora
ERR3862990	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862991	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862992	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862993	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862994	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862995	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862996	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862997	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine

ERR3862998	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3862999	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3863000	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
ERR3863002	EDL933	M9 minimal media	TUV93-0 or D-serine mutants in minimal media with or without D-serine
SRR10883929	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	Total RNA
SRR10883930	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	Total RNA
SRR10883931	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	Total RNA
SRR10883938	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	EHEC Term-seq
SRR10883939	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	EHEC Term-seq
SRR10883940	Sakai	MEM-Hepes supplemented with 0.1% glucose and 250 nM of Fe(NO3)3	EHEC Term-seq
SRR11026642	86-24	Low glucose DMEM	WT EHEC other set
SRR11026643	86-24	Low glucose DMEM	WT EHEC other set
SRR11026644	86-24	Low glucose DMEM	WT EHEC other set
SRR11026645	86-24	Low glucose DMEM with 10uM serotonin	WT EHEC with serotonin
SRR11026646	86-24	Low glucose DMEM with 10uM serotonin	WT EHEC with serotonin
SRR11026647	86-24	Low glucose DMEM with 10uM serotonin	WT EHEC with serotonin
SRR2637695	EDL933	LB	WT control

SRR2637696	EDL933	LB	WT control
SRR2637697	EDL933	LB	*hns* mutant EHEC
SRR2637698	EDL933	LB	*hns* mutant EHEC
SRR6029690	EDL933	M9 minimal media	Cra WT
SRR6029691	EDL933	M9 minimal media	Cra WT
SRR6029692	EDL933	M9 minimal media	CraY47D
SRR6029693	EDL933	M9 minimal media	CraY47D
SRR6029694	EDL933	M9 minimal media	CraY47F
SRR6029695	EDL933	M9 minimal media	CraY47F
SRR6869342	EDL933	Bovine small intestine content	EDL933 in bovine small intestine content
SRR6869343	EDL933	Bovine small intestine content	EDL933 in bovine small intestine content
SRR6869346	EDL933	Bovine small rumen content	EDL933 in bovine rumen content
SRR6869347	EDL933	Bovine small rumen content	EDL933 in bovine rumen content
SRR6869353	EDL933	M9 minimal media	EDL933 in M9 medium
SRR6869355	EDL933	M9 minimal media	EDL933 in M9 medium
SRR6869356	EDL933	M9 minimal media	EDL933 in M9 medium
SRR6869354	EDL933	Bovine small rectum content	EDL933 in bovine rectum content
SRR6869358	EDL933	Bovine small rectum content	EDL933 in bovine rectum content
SRR6869359	EDL933	Bovine small rectum content	EDL933 in bovine rectum content
SRR7782878	86-24	Low glucose DMEM	WT EHEC other set 2
SRR7782879	86-24	Low glucose DMEM	WT EHEC other set 2
SRR7782880	86-24	Low glucose DMEM	WT EHEC other set 2
SRR7782881	86-24	Low glucose DMEM	*tnaA* mutant EHEC

SRR7782882	86-24	Low glucose DMEM	*tnaA* mutant EHEC
SRR7782883	86-24	Low glucose DMEM	*tnaA* mutant EHEC
SRR7782884	86-24	Low glucose DMEM with 500uM indole	WT EHEC with indole
SRR7782885	86-24	Low glucose DMEM with 500uM indole	WT EHEC with indole
SRR7782886	86-24	Low glucose DMEM with 500uM indole	WT EHEC with indole
SRR7782887	86-24	Low glucose DMEM with 500uM indole	*tnaA* mutant with indole
SRR7782888	86-24	Low glucose DMEM with 500uM indole	*tnaA* mutant with indole
SRR7782889	86-24	Low glucose DMEM with 500uM indole	*tnaA* mutant with indole
SRR8271736	86-24	DMEM	WT EHEC
SRR8271737	86-24	DMEM	WT EHEC
SRR8271738	86-24	DMEM	WT EHEC
SRR8271739	86-24	DMEM	EHEC *dicF*1-4 deletion strain
SRR8271740	86-24	DMEM	EHEC *dicF*1-4 deletion strain
SRR8271741	86-24	DMEM	EHEC *dicF*1-4 deletion strain
SRR8485398	Sakai	CAMHB	WT control replicate
SRR8485399	Sakai	Cation-adjusted Muel*ler* Hinton Broth (CAMHB)	WT control replicate
SRR8485400	Sakai	Cation-adjusted Muel*ler* Hinton Broth (CAMHB)	WT control replicate
SRR8485401	Sakai	Cation-adjusted Muel*ler* Hinton Broth (CAMHB)	Treated replicate
SRR8485402	Sakai	Cation-adjusted Muel*ler* Hinton Broth (CAMHB)	Treated replicate
SRR8485403	Sakai	Cation-adjusted Muel*ler* Hinton Broth (CAMHB)	Treated replicate
1-1-d	EDL933	Saline	12h ringer
1-2-d	EDL933	Saline	12h ringer
1_2h_M9	FORC_041	M9 minimal media	M9 minimal media 2h incubation

2_2h_M9	FORC_041	M9 minimal media	M9 minimal media 2h incubation
3_2h_M9	FORC_041	M9 minimal media	M9 minimal media 2h incubation
1_4h_M9	FORC_041	M9 minimal media	M9 minimal media 4h incubation
2_4h_M9	FORC_041	M9 minimal media	M9 minimal media 4h incubation
3_4h_M9	FORC_041	M9 minimal media	M9 minimal media 4h incubation
1_2h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 2h incubation
2_2h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 2h incubation
3_2h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 2h incubation
1_4h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 4h incubation
2_4h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 4h incubation
3_4h_M9plusB	FORC_041	M9 minimal media with beef	M9 minimal media with beef 4h incubation

Minimum essential media, MEM; Dulbecco's modified Eagle medium, DMEM; Cation-adjusted Muel*ler* Hinton broth, CAMHB.

Table III-2. The trimmed and log-transformed (log_2 TMM+1) transcriptome data

	SRR number or experimental conditions						
Locus tag	**SRR6869353**	**SRR6869355**	**SRR6869356**	· · ·	**1_4h_M9plusB**	**2_4h_M9plusB**	**3_4h_M9plusB**
Z0001	0.108492	-0.01469	-0.0938	· · ·	-0.4296	-0.67172	-0.42913
Z0002	0.316487	-0.20536	-0.11112	· · ·	-1.60319	-1.10017	-1.79623
Z0003	0.332196	-0.08157	-0.25062	· · ·	-1.3048	-1.09442	-1.60055
Z0004	0.288648	-0.14544	-0.1432	· · ·	-2.26621	-2.60746	-2.58908
Z0005	0.200272	-0.15113	-0.04914	· · ·	-0.37682	-1.20077	-1.05291
Z0006	-0.24585	0.069727	0.176119	· · ·	0.392789	-0.0056	0.151563
Z0007	0.270997	-0.50143	0.230433	· · ·	2.653517	1.922387	2.290606
Z0008	-0.13572	0.075031	0.060689	· · ·	0.008381	0.857904	0.384026
· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·
Z_L7092	0.102952	-0.06082	-0.04213	· · ·	-1.54774	-2.01974	-1.69791
Z_L7093	-0.31133	0.459338	-0.14801	· · ·	-2.96977	-3.7029	-3.24386
Z_L7094	-0.49259	0.718731	-0.22614	· · ·	-1.2687	-1.97407	-1.30009
Z_L7095	0.051658	0.026144	-0.0778	· · ·	-9.21149	-9.21149	-9.21149
Z_L7096	-0.4013	0.790972	-0.38968	· · ·	-0.49984	-0.83448	-0.56045
Z_L7097	0.163671	0.256404	-0.42008	· · ·	-2.50759	-2.50759	-2.50759
Z_L7098	-0.05854	0.27696	-0.21842	· · ·	1.885164	1.163424	1.406993
Z_L7099	-0.18225	0.155324	0.026929	· · ·	-1.53669	-2.63994	-1.65653
Z_L7100	-0.05432	0.072751	-0.01843	· · ·	-7.02786	-7.02786	-7.02786
Z_L7101	-0.50057	0.337264	0.163308	· · ·	0.397737	0.028382	-0.15296

Transcriptome data from only partial experimental conditions are presented because whole data are too large to be displayed in the table.

Full data can be found at https://github.com/hanhyeok/EHEC_ICA.

III-2-2. Identification of the modulons by using ICA

ICA was applied to the trimmed transcriptome data as previously described (Sastry *et al.*, 2019). Briefly, all trimmed transcriptome data were centered using the mean read counts of the transcriptome data of the EHEC EDL933 grown in M9 minimal medium. ICA with random seed was executed 256 times to construct the independent gene components from the trimmed transcriptome data using the FastICA algorithm from the Scikit-learn Python package (Varoquaux *et al.*, 2015). The D'Agostino K^2 test, which measures the skewness and kurtosis of distribution, was performed on the gene coefficients of the element genes in the independent components to select the co-regulated genes of the components (D'agostino *et al.*, 1990). The element gene with the greatest absolute coefficient in each independent gene component was repeatedly removed, and the D'Agostino K^2 test statistic was calculated for each removal. If the test statistic dropped below a cutoff, the removed genes were defined as the co-regulated genes of the independent component.

To determine the K^2 test statistic cutoff, a two-sided Fisher's exact test was performed between the previously known regulons of the *E. coli* regulators and the top 25 element genes of the independent gene components (Gama-Castro *et al.*, 2011; Fang *et al.*, 2017; Sastry *et al.*, 2019). Among the regulators, the regulator with the lowest *P* value was linked to each independent gene component. Then, the F1 scores were calculated between the regulons of the component-linked regulators and the co-regulated genes of the independent gene components selected based on the K^2 test statistic cutoff varying from 1,500 to 2,500. Because the average of calculated F1

scores showed a maximum value at the K^2 test statistic cutoff of 1,800 (Fig. III-1D), the statistic cutoff was used to define the modulons. The independent components with less than 5 co-regulated genes were discarded and thus, the 64 modulons were identified from the 85 independent components. The 64 modulons were named after their related regulator or biological function (e.g., H-NS or LEE). Detailed information of the modulons, such as the related TF or biological function, the co-regulated genes, and the gene coefficients, was available in Table III-3.

Table III-3. The trimmed and log-transformed (log_2 TMM+1) transcriptome data

Component number	Locus tag	Coefficient	Annotated gene	Description	Summary
0	Z0070	0.066234	*araA*	L-arabinose isomerase	carbon compounds
0	Z0072	0.068409	*araB*	ribulokinase	carbon compounds
0	Z1675	0.068349	*csgB*	curlin, minor subunit	glycoprotein
0	Z2211	0.07351		putative anaerobic sulfatase maturation enzyme	chaperoning, repair
0	Z2712	-0.06764	*sufA*	iron-sulfur cluster insertion protein SufA	incorporation of metal ions
0	Z2953	0.072843	*araG*	arabinose ABC transporter ATP binding subunit	carbon compounds
0	Z2954	0.067599	*araF*	arabinose ABC transporter periplasmic binding protein	carbon compounds
0	Z3714	0.068795	*eutD*	phosphate acetyltransferase EutD	carbon utilization
0	Z4209	0.086803		putative carbamoyl transferase YgeW	
0	Z4210	0.066739	*ygeX*	2,3-diaminopropionate ammonia-lyase	metabolism
0	Z4211	0.086852	*ygeY*	putative peptidase YgeY	
0	Z4466	0.065174	*tdcE*	activated 2-ketobutyrate formatelyase/pyruvate formatelyase 4	amino acids
0	Z4467	0.08556	*tdcD*	propionate kinase	threonine catabolism
0	Z4468	0.091815	*tdcC*	threonine/serine: H^+ symporter	threonine
0	Z4469	0.106961	*tdcB*	catabolic threonine dehydratase	amino acids
0	Z5203	0.06887	*tnaA*	tryptophanase	amino acids
0	Z5285	-0.07788	*ilvC*	ketol-acid reductor isomerase ($NADP^+$)	isoleucine/valine
...	...	...	...	...	...
84	Z1371	0.076311		CP-933M	
84	Z_L7101	-0.0676	*flmB*		

84	Z2543	-0.06706	*yciO*		
84	Z2380	0.064439		tRNA-Arg	
84	Z2381	0.06375			
84	Z3626	-0.06302			
84	Z5862	0.061871	*yjgJ*		
84	Z3086	-0.06107		CP-933U	
84	Z3048	0.060012	*yodC*		
84	Z6038	0.05993		CP-933P	
84	Z2249	-0.05955	*nhoA*		
84	Z4536	-0.05921		tRNA-Leu	
84	Z1917	0.05894		CP-933X	
84	Z2127	-0.05841		CP-933O	
84	Z2040	0.057623		CP-933O	
84	Z2360	0.056771		CP-933R	
84	Z3400	0.056668	*yeiS*		
84	Z4337	-0.05623			
84	Z0906	0.055886	*tolR*		membrane
84	Z1851	0.055726		CP-933C	
84	Z6045	0.05429		CP-933P	
84	Z1676	0.053907	*csgA*		fimbriae, pili
84	Z2181	0.053082			

Data from only partial modulons are presented because the entire data of whole modulons are too large to be displayed in the table.

Full data can be found at https://github.com/hanhyeok/EHEC_ICA.

III-2-3. Calculation of cumulative explained variance (CEV) for principal component analysis (PCA) and ICA

The PCA of the trimmed transcriptome data was performed with the Scikit-learn Python package (Varoquaux *et al.*, 2015). The CEV for the PCA results was calculated by sequentially adding the explained variance ratios of the principal components using Scikit-learn and NumPy Python packages (Varoquaux *et al.*, 2015; Harris *et al.*, 2020). The CEV for the ICA results was calculated as previously described in EEGLAB (Delorme and Makeig, 2004). The Matplotlib Python package was used to visualize the CEV for the PCA and the ICA results (Hunter, 2007).

III-2-4. The correlation analyses of the expression levels of the genes or the activities of the modulons

The expression levels of the genes and the activities of the modulons were obtained from Table III-2 and III-4, respectively. The Pearson correlation analyses between the expression levels of the genes and the activities of the modulons were performed with the SciPy Python package (Virtanen *et al.*, 2020). The Pearson correlations between the expressions of the different genes were performed with the Pandas Python package (McKinney, 2010). The Matplotlib Python package was used to visualize

Table III-4. The activities of the modulons under different experimental conditions transcriptome data

	SRR number or experimental conditions						
Component number	**SRR6869353**	**SRR6869355**	**SRR6869356**	**⋯**	**1_4h_M9plusB**	**2_4h_M9plusB**	**3_4h_M9plusB**
0	0.223	1.667	-1.889	⋯	21.267	18.100	20.193
1	2.498	-1.544	-0.954	⋯	6.379	2.074	5.177
2	0.244	-0.043	-0.201	⋯	-1.193	-3.444	-2.615
3	-2.061	3.350	-1.289	⋯	-16.112	-10.530	-12.119
4	-0.234	-0.099	0.333	⋯	-1.244	0.977	0.199
5	0.687	-0.881	0.194	⋯	-0.420	0.294	0.862
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
80	-0.541	1.364	-0.823	⋯	0.020	-5.057	-2.121
81	-0.013	1.122	-1.109	⋯	-1.122	-2.013	-1.058
82	1.676	-1.647	-0.029	⋯	6.440	3.520	4.982
83	-0.269	1.201	-0.931	⋯	4.175	3.465	3.973
84	-0.503	0.476	0.027	⋯	3.768	4.069	3.695

Data from only partial modulons are presented because the entire data of whole modulons are too large to be displayed in the table.

Full data can be found at https://github.com/hanhyeok/EHEC_ICA.

III-2-5. Searching for the Ler binding site of the Z0395 gene

The binding motif of Ler was discovered from the specific binding sequences of Ler, which were previously reported by Hiroyuki *et al.* (Abe *et al.*, 2008), by using the Multiple Expectation maximizations for Motif Elicitation (MEME) (Bailey *et al.*, 2006). The Ler binding site was predicted *in silico* by searching the upstream sequences of the Z0395 gene by using the Find Individual Motif Occurrences (FIMO) (Grant *et al.*, 2011).

III-2-6. Strains, plasmids, and culture conditions

All the strains and plasmids used in this study are listed in Table III-5**.** Unless otherwise noted, the *E. coli* strains were grown aerobically in the Luria-Bertani (LB) medium at 37°C. *E. coli* DH5α was used as a cloning host, and EHEC EDL933 was used as the WT. The pCas and pTargetF plasmids required for mutant construction of *E. coli* were obtained from Addgene (plasmid #62225 and #62226) (Jiang *et al.*, 2015).

Table III-5. Bacterial strains and plasmids used in this study

Strain or plasmid	Relevant characteristics[a]	Reference or source
Bacterial strains		
E. coli		
DH5α	*supE44 ΔlacU169* (*Φ80 lacZ Δ*M15) *hsdR17 recA1 endA1 gyrA96 thi-1 relAI*	Laboratory collection
EDL933	Wild-type; clinical isolate; virulence	Laboratory collection
HH101	EDL933 with Δ*ler*	This study
Plasmids		
pCas	*repA101*(Ts) P_{cas}-*cas9* P_{araB}-*Red lacI*^{q} P_{trc}-*sgRNA-pMB1*; Km^{r}	(Jiang *et al.*, 2015)
pTargetF	*pMB1 aadA sgRNA-cadA*; Spc^{r}	(Jiang *et al.*, 2015)
pTargetF-ler	pTargetF with *sgRNA-ler*; Spc^{r}	This study

[a]Km^{r}, kanamycin-resistant; Spc^{r}, spectinomycin-resistant.

III-2-7. Generation of a *ler* deletion mutant

The *ler* (Z5140) was inactivated by deletion (207 bp of 390 bp) of the coding region using the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 system as previously described (Jiang *et al.*, 2015). Briefly, two amplicons designed to carry homologous arms with 5′- and 3′-flanking regions of *ler* were amplified by PCR using LER-F1-F and -R or LER-F2-F and -R pairs of primers (Table III-6). Both amplicons were fused into donor DNA by overlap extension PCR using the primer pairs of LER-F1-F and LER-F2-R. Replacing the N_{20} of pTargetF to target *ler* was performed using the Site-Directed Mutagenesis Kit (NEB, Beverly, MA) according to the manufacturer's protocols. The N_{20} replaced pTargetF targeting *ler* was designated as pTargetF-*ler* (Table III-5). The EDL933 electrocompetent cells harboring pCas were prepared as previously described (Sharan *et al.*, 2009). For genome editing, 400 ng of donor DNA and 100 ng of pTargetF-*ler* were co-electroporated into the EDL933 electrocompetent cells. The construction of the *ler* deletion mutant was confirmed by PCR.

Table III-6. Oligonucleotides used in this study

Oligonucleotide	Oligonucleotide sequence (5' →3')[a]	Use
For mutant construction		
LER-F1-F	AAAACATTTGCGGCTTCTTT	Deletion of *ler* ORF
LER-F1-R	CTACAGCAGGAAGCAGAAGCACTGTTGAATGGAATGAAGAAAGAAGATTT	
LER-F2-F	ATTCAACAGTGCTTCTGCTTCCTGCTGTAGAACTGCAATTTGCTCTATAA	
LER-F2-R	CAGGAAGGACCAACAATTAATCA	
N20-LER-F	TTCTTCATTGGTTTTAGAGCTAGAAATAGC	Replacement of N_{20} of pTargetF
N20-LER-R	GGGCAGACCTACTAGTATTATACCTAGGAC	
For qRT-PCR		
GAPDH-qRT-F	AGGTCTGATGACCACCGTTC	Quantification of the *gapA* expression
GAPDH-qRT-F	AACGGTCAGGTCAACTACGG	
Z0395-qRT-F	AAAGCCAGTCTCCTTCAACTC	Quantification of the Z0395 gene expression
Z0395-qRT-R	CTCGACAACACATCCTCTTCTT	
STX2A-qRT-F	GAACGTTCCGGAATGCAAA	Quantification of the *stx2a* expression
STX2A-qRT-R	CCATTAACGCCAGATATGATGA	
THIB-qRT-F	CGAAGGCGAAGTAGCCATAA	Quantification of the *thiB* expression
THIB-qRT-R	GTTAGACGCCGCCAGTAAA	
THIC-qRT-F	CCGACGTGAAGTGGTCATAG	Quantification of the *thiC* expression
THIC-qRT-R	GCAATATGACCGAGGAGTTAGAG	
CUSC-qRT-F	CGCTTAAAGAACATGAGCGAAG	Quantification of the *cusC* expression
CUSC-qRT-R	TAGCTTTGCGCGACATTA	

[a]Regions of oligonucleotides not complementary to the corresponding genes are underlined.

III-2-8. Quantitative reverse transcription-PCR (qRT-PCR)

The total RNA of the EDL933 strains grown under various conditions were isolated to determine the relative transcript levels of genes of interest by qRT-PCR. In detail, to determine the relative transcript levels of the Z0395 gene, the EHEC strains were grown in low-glucose Dulbecco's modified Eagle's medium (Merck, Darmstadt, Germany) at 37℃ to an A_{600} of 1.0. To determine the relative transcript levels of *thiB*, *thiC*, and *stx2a*, the EHEC strains were grown in M9 minimal medium with or without thiamine at 37℃ to an A_{600} of 0.75. Finally, to determine the relative transcript levels of *cusC* and *stx2a*, the EHEC strains were grown in LB medium with different levels of $CuSO_4$ at 37℃ to an A_{600} of 1.0. The total RNAs of the strains were isolated using the RNeasy mini kit (Qiagen, Valencia, CA, USA). For qRT-PCR, the concentrations of the total RNAs were measured by using a NanoDrop One spectrophotometer (Thermo Scientific, Waltham, MA, USA), and cDNA was synthesized from 100 ng of total RNA by using iScript cDNA synthesis kit (Bio-Rad, Hercules, CA, USA). Real-time PCR amplification of the cDNA was performed by using CFX96 real-time PCR detection system (Bio-Rad) with specific primer pairs (Table III-6) as described previously (Jang *et al.*, 2017). The relative transcript levels of the genes were calculated by using the transcript levels of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as the internal reference for normalization (Kijewski *et al.*, 2020).

III-3. Results

III-3-1. The modulons containing target genes of several TRNs of EHEC are identified by using ICA

ICA, a machine learning-based decomposition method, was used to decompose the large-scale transcriptome data of EHEC into the modulons containing the target genes of several TRNs. For this purpose, the trimmed 88 transcriptome data of EHEC ($R^2 \geq 0.9$ between biological replicates) (Fig. III-1A, Table III-1 and III-2) were decomposed into the 85 independent gene components (Table III-3). ICA was also used to calculate the overall expression levels of the decomposed 85 components: the activities of the components in a specific condition. The activities of the 85 independent components (Table III-4) successfully explained 83% of the total expression variance of the 88 transcriptome data (Fig. III-1B), validating that ICA properly decomposed the transcriptome data of EHEC into the independent gene components.

The 85 independent components contain the element genes with varied gene coefficients that represent the degree of the regulatory effect on the expressions of the genes. The element genes with a positive or negative gene coefficient indicate that their expressions are proportionally or inversely regulated along with the activities of the independent component, respectively. Unless otherwise noted, gene coefficient signs of element genes in an independent component are positive. Most

of the gene coefficients of element genes in an independent component were distributed close to 0 (Fig. III-1C), indicating that the expressions of only a few element genes significantly rely on the activities of an independent component. The distribution of the gene coefficients was reexamined by the statistical analysis, D'Agostino K^2 test (D'agostino *et al.*, 1990; Sastry *et al.*, 2019), to select only the genes with the coefficients far away from 0. As a result, the element genes with the gene coefficients over a cutoff, D'Agostino K^2 test statistic 1,800 (Fig. III-1D), were selected as the co-regulated genes of an independent component and defined as the modulons (see *Materials and Methods* for details on the selection process). Consequently, a total of 64 modulons were identified from the 85 independent components. The 64 modulons with detailed information are presented in Table III-3.

III-3-2. The LEE and Stx modulons contain the Ler regulon and the Stx prophage genes, respectively

The modulons mainly consisting of the LEE and the Stx prophage genes encoding the major virulence factors of EHEC were defined as the LEE and the Stx modulon, respectively, and were further investigated to confirm whether the modulons adequately contain the sets of co-regulated genes. The LEE modulon contains 44 genes, of which 39 were the LEE genes (Fig. III-2A). Also, the activities of the LEE modulon were strongly correlated with the expression levels of *ler* (Pearson $R = 0.79$,

$p < 10^{-10}$) (Fig. III-2B), indicating that the LEE modulon primarily consisted of the Ler regulon. Furthermore, the LEE modulon contained *lpxR*, *nleA*, *stcE*, and *etpC*, which were not located in the LEE but known as the Ler regulon (Fig. III-2A) (Grys *et al.*, 2005; Tobe *et al.*, 2006; Roe *et al.*, 2007; Ogawa *et al.*, 2018). The activities of the LEE modulon were also highly correlated with the expression levels of these genes (Pearson $R > 0.5$, $p < 10^{-5}$) (Fig. III-2C), indicating that the modulon properly contained the genes located separately but co-regulated by Ler. Similarly, the Stx modulon contains the CP-933V and BP-933W prophage genes that include *stx1* and *stx2,* respectively (Fig. III-2D). The activities of the Stx modulon were also highly correlated with the expression levels of the antiterminator *Qs* (Pearson $R > 0.5$, $p < 10^{-5}$) (Fig. III-2 E and F), indicating that the Stx modulons mainly consisted of the Stx prophage genes co-related by the antiterminator *Qs*. Consequently, these results validated that the modulons, the independent sets of co-regulated genes, were appropriately identified from the large-scale transcriptome data by using ICA.

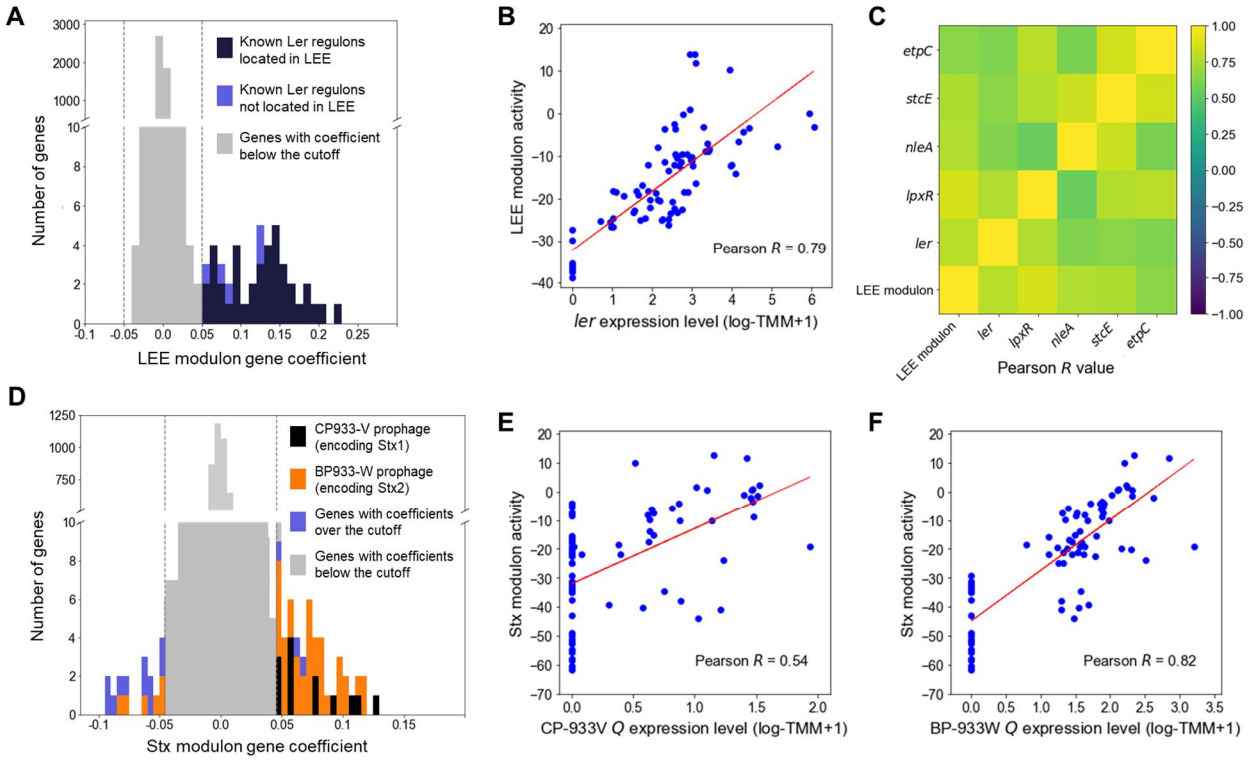

Figure III-2. Validation of the LEE and Stx modulons. (A, D) Histograms of the gene coefficients of the element genes in the LEE (A) and the Stx modulons (D). The dotted lines in the boxes show the cutoff values of the gene coefficients in each modulon. The colors of the bars represent

the classifications of the genes as indicated in the plots. (B, E, and F) The scatter plots of the activities of the LEE modulon and the expression levels of the *ler* (B), the activities of the Stx modulon and the expression levels of the CP-933V antiterminator *Q* (E), and the activities of the Stx modulon and the expression levels of the BP-933W antiterminator *Q* (F). The Pearson *R* values between the activities of the modulons and the expression levels of their related TF are denoted in the boxes. Each dot of the plots represents a single biological replicate. Red lines represent the regression lines of the plots. (C) Ordered correlation matrix. Colors indicate the Pearson *R* values between the activities of the LEE modulon and the expression levels of the Ler regulon that are not located in the LEE. Yellow and indigo represent the strongest positive (+1) and negative (-1) correlation, respectively.

III-3-3. The LEE modulon contains the Z0395 gene as a novel member of the Ler regulon

The element genes of the LEE modulon were further investigated to analyze the target genes of the Ler TRNs encoding the major virulence factor of EHEC. The LEE modulon included a hypothetical Z0395 gene, which is not located in the LEE and is not known as the Ler regulon. Since most of the genes in the LEE modulon were the Ler regulon (Fig. III-3A), it was possible that the Z0395 gene is also a member of the Ler regulon. To examine the possibility, the relationship between the expressions of the Z0395 gene and *ler* was analyzed. The expressions of the Z0395 gene and *ler* were positively correlated (Pearson $R = 0.33$, $p < 0.05$) (Fig. III-3A). Thus, to further verify the effect of Ler on the the transcript levels of the Z0395 gene in the WT and the *ler* deletion mutant (Δ*ler*) were compared. The transcript level of the Z0395 gene was greatly reduced in Δ*ler* (Fig. III-3B), confirming that Ler activates the Z0395 gene expression at the transcription level. To examine whether Ler directly binds to the probable promoter region of the Z0395 gene, the upstream region of the gene was scanned *in silico* with the binding motif of Ler. The motif-based sequence analysis predicted one Ler binding sequence located in the -212 to -201 region from the open reading frame (ORF) of the Z0395 gene ($p < 10^{-5}$) (Fig. III-3 C and D). Taken together, these results indicated that Ler regulates the expression of the Z0395 gene by directly binding to its upstream region, supporting that the Z0395 gene in the LEE modulon is a novel member of the Ler regulon.

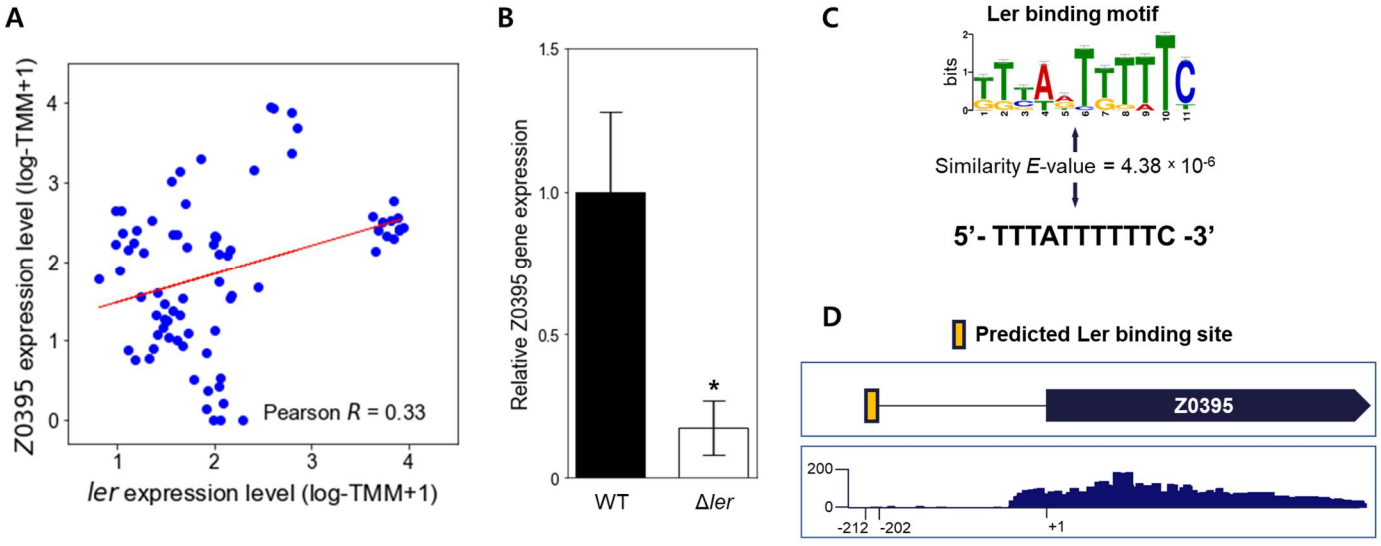

Figure III-3. The Z0395 gene is a member of the Ler regulon. (A) The scatter plot of the expression levels of the Z0395 gene and *ler*. Each dot of the plot represents a single biological replicate. Red line represents the regression line of the plot. The Pearson *R* value between the expression levels of the Z0395 gene and *ler* is denoted in the box. (B) The relative expression levels of the Z0395 gene in the WT and *ler* deletion mutant. The levels of the Z0395 gene transcripts were determined by qRT-PCR, and the Z0395 gene transcript levels in the WT were set to 1. Error bars represent the SD from four independent experiments. Statistical significance was determined by the Student's *t* test (*, $p < 0.05$). WT, EDL933;

Δ*ler*, *ler* deletion mutant. (C) The Ler binding motif depicted in the logo and the Ler binding sequence predicted *in silico* found at the Z0395 gene upstream region. The height of the letters in the logo represents the information contents of the position in bits. The similarity between the Ler binding motif (top) and the predicted binding sequence (bottom) are denoted as *E*-value. (D) Location of the Ler binding sequence *in silico* predicted in the Z0395 gene upstream region. The Ler binding sequence is located from -212 to -202 region of the Z0395 ORF, represented as a yellow box. The bellow box represents the coverage plot of the reads mapped to the Z0395 gene. The transcriptome data of EHEC EDL933 grown in M9 minimal medium were used to generate the plot. The *y*-axis represents the normalized number of reads per base. The average number of reads of the biological triplicates are shown in the plot.

III-3-4. The Stx modulon contains the *thi* and *cus* locus genes in addition to the Stx prophages

The element genes composing the Stx modulon were also further investigated. The Stx modulon contained the *thi* locus genes *thiBP* and *thiCEFGH* and the *cus* locus genes *cusCFBA*, which are not located in the Stx prophages (Fig. III-4A). These genes have negative gene coefficients in the Stx modulon, unlike the Stx prophage genes with positive gene coefficients (Fig. III-4A), indicating that the expressions of the *thi* and *cus* locus genes decrease as the activities of the Stx modulon increase. In accordance with this, the expression levels of *thiB, thiC*, and *cusC* were negatively correlated with the activities of the Stx modulon, with Pearson *R* -0.57 ($p < 10^{-5}$), -0.72 ($p < 10^{-10}$), and -0.61 ($p < 10^{-5}$), respectively (Fig. III-4 B, C, and D). The negative relationship were also verified with the Spearman's rank correlation coefficients and Kendall τ correlation coefficients (Table III-7). To verify whether the relationship is observed due to certain transcriptome data from an experimental condition, we recalculated the correlation coefficients by randomly excluding the transcriptome data from one experimental condition. As a result, it was confirmed that there was no significant difference between the coefficients calculated from the entire transcriptome data and the randomly excluded transcriptome data (Fig. III-5 A and B). The result indicated that the correlation between Stx modulon activity and expression levels of *thi* and *cus* locus genes is not dependent on the transcriptome data from certain experimental condition. The

negative relationship was further verified by the correlation analyses between the expression levels of *thiB, thiC*, and *cusC,* and those of *stx2a* (Pearson $R < -0.5$, $p < 10^{-8}$) (Fig. III-4E), indicating that the expression patterns of the *thi* and *cus* locus genes were contrary to those of the Stx prophage genes.

Since the expressions of the *thi* and *cus* locus genes are regulated by the levels of thiamine and copper ions, respectively (Vander Horn *et al.*, 1993; Webb *et al.*, 1998; Miranda-Rios *et al.*, 2001; Delmar *et al.*, 2015), the effect of the nutrients on the expression of *stx2a* was examined. Interestingly, the presence of thiamine decreased significantly the transcription of *thiB* and *thiC,* but increased that of *stx2a* (Fig. III-4F). Copper ions also increased the transcription of *cusC,* but decreased that of *stx2a* in a dose-dependent manner (Fig. III-4G). Consequently, the combined results revealed that the Stx modulon includes the *thi* and *cus* locus genes in addition to Stx prophage genes, which are regulated by the levels of thiamine and copper ions.

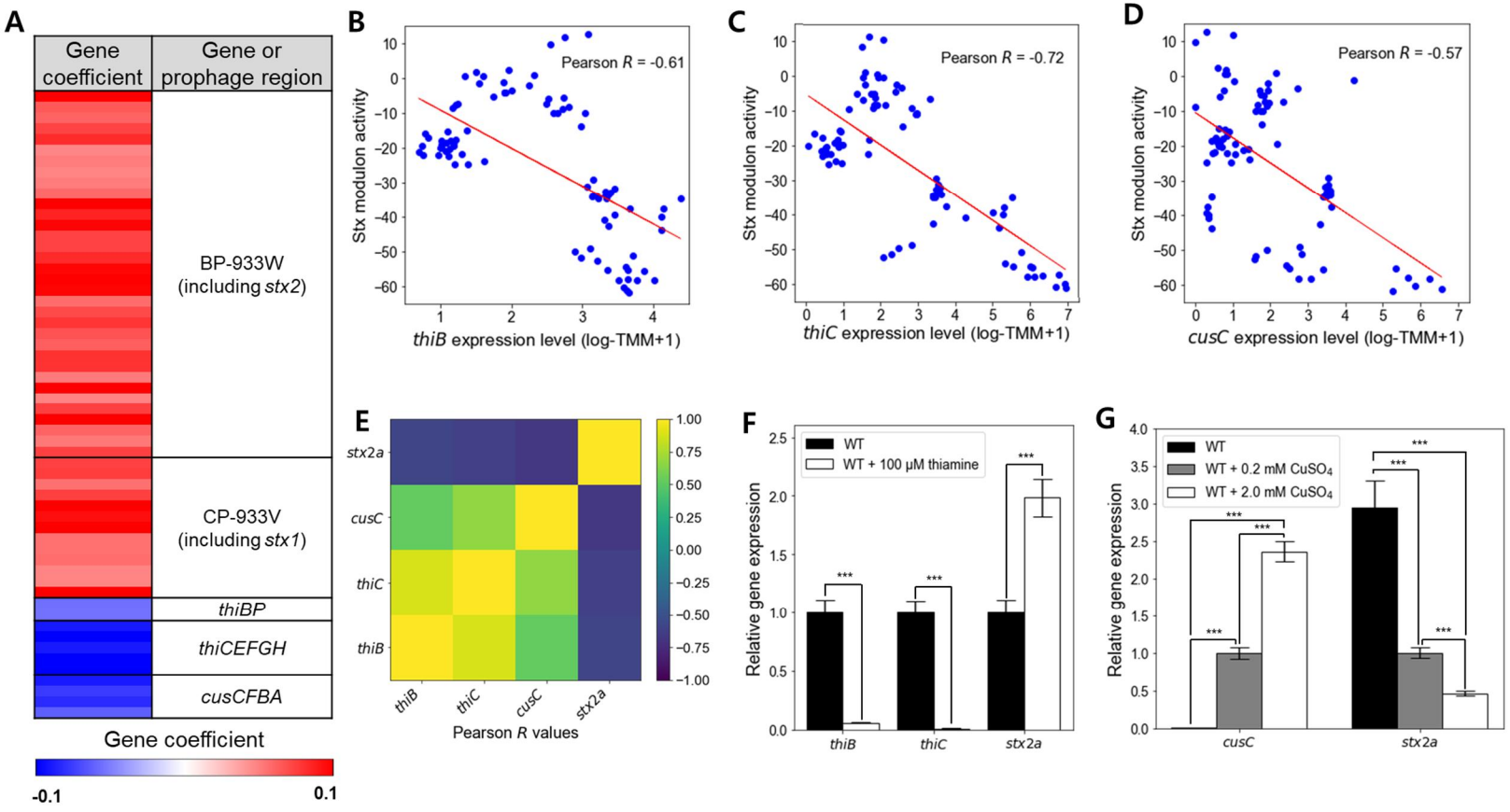

Figure III-4. The contrary expression patterns of the *thi* and *cus* locus genes to those of the Stx prophage genes. (A) Heatmap for the gene

coefficients in the Stx modulon. Red and blue represent the high (+0.1) and low (-0.1) gene coefficient, respectively. (B, C, and D) The scatter plots of the activities of the Stx modulon and the expression levels of *thiB* (B), *thiC* (C), and *cusC* (D). Each dot of the plots represents a single biological replicate. Red lines represent the regression lines of the plots. The Pearson *R* values between the activities of the Stx modulon and the expression levels of each gene of the plot are denoted in the boxes. (E) Ordered correlation matrix. Colors indicate the Pearson *R* values between the expression levels of *thiB*, *thiC*, *cusC,* and *stx2a,* as indicated. Yellow and indigo represent the strongest positive (+1) and negative (-1) correlation, respectively. (F and G) The relative expression levels of genes of interest in the WT grown under the different levels of thiamine and copper ions. The transcript levels of *thiB*, *thiC*, and *stx2a* in the WT with or without thiamine were determined by qRT-PCR, and the transcript levels of each gene in the WT were set to 1 (F). The transcript levels of *cusC* and *stx2a* in the WT with the different levels of $CuSO_4$ were also determined by qRT-PCR, and the transcript levels of each gene in the WT with 0.2 mM $CuSO_4$ were set to 1 (G). Error bars represent the SD from four independent experiments. Statistical significance was determined by the Student's *t* test (***, $p < 10^{-3}$). WT, EDL933.

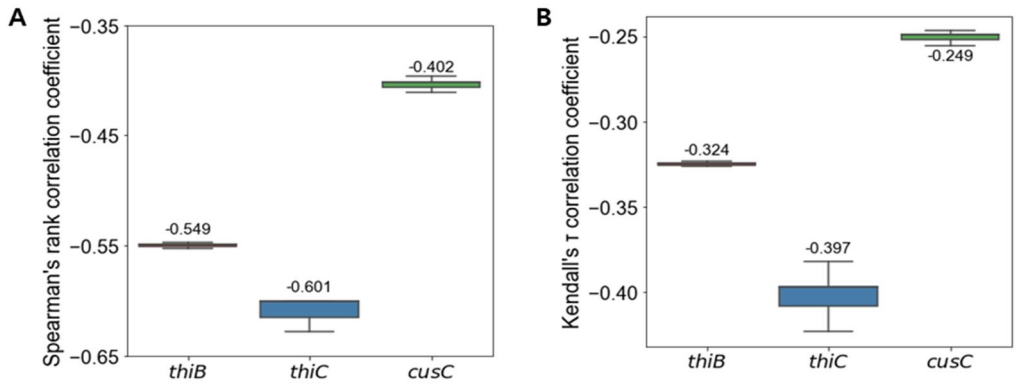

Figure III-5. The box plots of Spearman's rank and Kendall's τ correlation coefficients calculated from the transcriptome data excluding one experimental condition.

Table III-7. Spearman and Kendall τ correlation coefficients between the Stx modulon activity and expression levels of *thiB*, *thiC*, and *cusC*

	Spearman (P)	Kendall τ (P)
thiB	-0.549 ($3.1*10^{-8}$)	-0.324 ($7.6*10^{-6}$)
thiC	-0.601 ($6.0*10^{-10}$)	-0.397 ($4.3*10^{-8}$)
cusC	-0.402 ($1.0*10^{-4}$)	-0.249 ($5.9*10^{-4}$)

III-3-5. The modulons enhance clustering the genes co-regulated regardless of the growth conditions

The element genes of the modulons are expected to be co-regulated under the various growth conditions. To verify this, it was investigated whether the expressions of the element genes in a modulon are altered together. The activities of the modulons were obtained from the transcriptome data of EHEC under different experimental conditions (Fig. III-5). Among them, the significantly changed activities of the LEE modulon were observed from the transcriptome data of the wild-type (WT) and *tna* deletion mutant (Δ*tna*) in the presence or absence of 500 μM indole. In Δ*tna* imitating EHEC grown without indole, the activities of the LEE modulon increased significantly ($p < 10^{-5}$) (Fig. III-6A). Accordingly, the expressions of the LEE genes, such as *escE*, *escJ*, *cesL*, *sepL*, and *tir,* increased significantly (Fig. III-6B). The addition of 500 μM indole significantly decreased the activities of the LEE modulon ($p < 10^{-5}$) (Fig. III-6A) and thereby decreased the expressions of the LEE genes (Fig. III-6B). Interestingly, the changed activities of the LEE modulon altered the expressions of the non-LEE located hypothetical gene Z0395 (Table III-3), the novel element gene of the LEE modulon (Fig. III-3A), in addition to the LEE genes (Fig. III-6B). These results indicated that the LEE modulon, as an example of the EHEC modulons, indeed enhanced the clustering of the co-regulated genes regardless of the growth conditions.

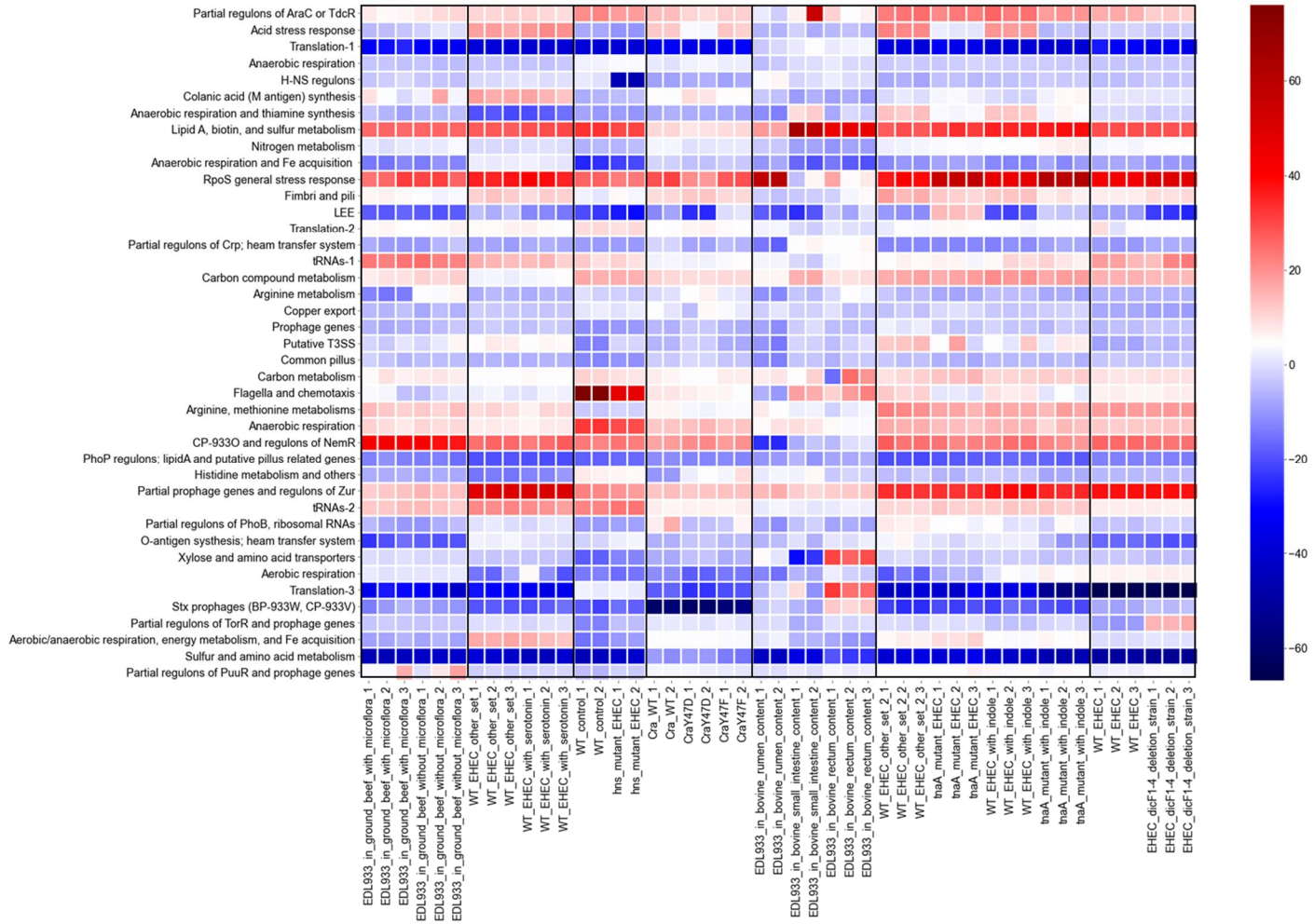
Partial regulons of AraC or TdcR
Acid stress response
Translation-1
Anaerobic respiration
H-NS regulons
Colanic acid (M antigen) synthesis
Anaerobic respiration and thiamine synthesis
Lipid A, biotin, and sulfur metabolism
Nitrogen metabolism
Anaerobic respiration and Fe acquisition
RpoS general stress response
Fimbri and pili
LEE
Translation-2
Partial regulons of Crp; heam transfer system
tRNAs-1
Carbon compound metabolism
Arginine metabolism
Copper export
Prophage genes
Putative T3SS
Common pillus
Carbon metabolism
Flagella and chemotaxis
Arginine, methionine metabolisms
Anaerobic respiration
CP-933O and regulons of NemR
PhoP regulons; lipidA and putative pillus related genes
Histidine metabolism and others
Partial prophage genes and regulons of Zur
tRNAs-2
Partial regulons of PhoB, ribosomal RNAs
O-antigen systhesis; heam transfer system
Xylose and amino acid transporters
Aerobic respiration
Translation-3
Stx prophages (BP-933W, CP-933V)
Partial regulons of TorR and prophage genes
Aerobic/anaerobic respiration, energy metabolism, and Fe acquisition
Sulfur and amino acid metabolism
Partial regulons of PuuR and prophage genes
EDL933_in_ground_beef_with_microflora_1
EDL933_in_ground_beef_with_microflora_2
EDL933_in_ground_beef_with_microflora_3
EDL933_in_ground_beef_without_microflora_1
EDL933_in_ground_beef_without_microflora_2
EDL933_in_ground_beef_without_microflora_3
WT_EHEC_other_set_1
WT_EHEC_other_set_2
WT_EHEC_other_set_3
WT_EHEC_with_serotonin_1
WT_EHEC_with_serotonin_2
WT_EHEC_with_serotonin_3
WT_control_1
WT_control_2
hns_mutant_EHEC_1
hns_mutant_EHEC_2
Cra_WT_1
Cra_WT_2
CraY47D_1
CraY47D_2
CraY47F_1
CraY47F_2
EDL933_in_bovine_rumen_content_1
EDL933_in_bovine_rumen_content_2
EDL933_in_bovine_small_intestine_content_1
EDL933_in_bovine_small_intestine_content_2
EDL933_in_bovine_rectum_content_1
EDL933_in_bovine_rectum_content_2
EDL933_in_bovine_rectum_content_3
WT_EHEC_other_set_2_1
WT_EHEC_other_set_2_2
WT_EHEC_other_set_2_3
tnaA_mutant_EHEC_1
tnaA_mutant_EHEC_2
tnaA_mutant_EHEC_3
WT_EHEC_with_indole_1
WT_EHEC_with_indole_2
WT_EHEC_with_indole_3
tnaA_mutant_with_indole_1
tnaA_mutant_with_indole_2
tnaA_mutant_with_indole_3
WT_EHEC_1
WT_EHEC_2
WT_EHEC_3
EHEC_dicF1-4_deletion_strain_1
EHEC_dicF1-4_deletion_strain_2
EHEC_dicF1-4_deletion_strain_3
60
40
20
0
−20
−40
−60

Figure III-6. Heatmap for the changed activities of the modulons obtained from the transcriptome data of EHEC under different experimental conditions. Only the modulons with known related TFs or biological functions were shown. The numbers on the labels indicate a distinct single biological replicate. Detailed experimental conditions of the transcriptome data can be found at Table III-1. Detailed activities of the modulons of the transcriptome data can be found at Table III-4. Red and blue represent the high and low activity of the modulon, respectively.

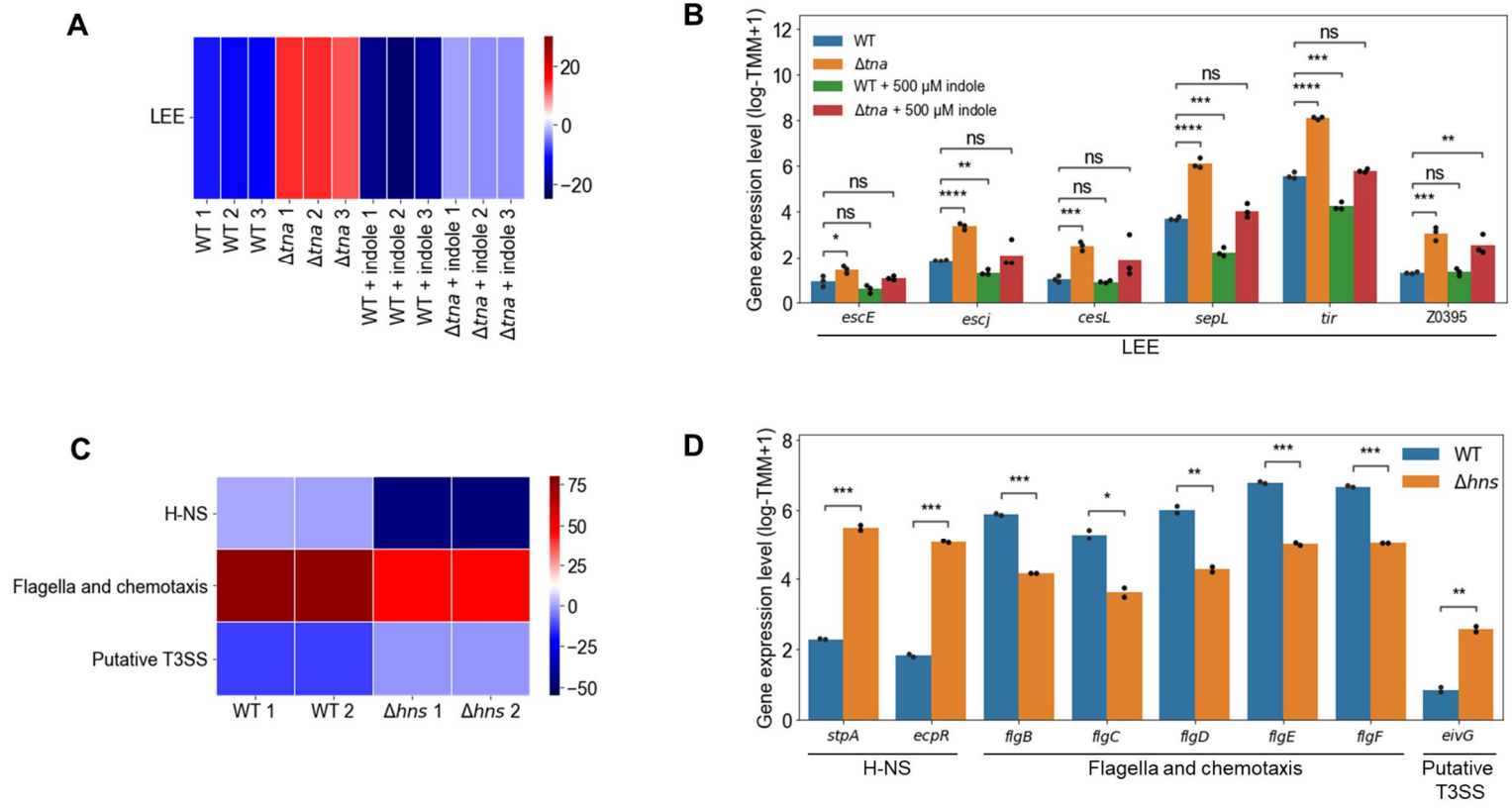

Figure III-7. The changed activities of the modulons obtained from the transcriptome data of EHEC EDL933 and its isogenic mutants.

(A, C) Heatmap for the changed activities of the modulons obtained from the transcriptome data of the WT and Δ*tna* in the presence or absence

of 500 μM indole (A), and the WT and Δ*hns* (C). The numbers on the bottom labels indicate a distinct single biological replicate. Red and blue represent the high and low activity of the modulon, respectively. WT, EDL933; Δ*tna, tna* deletion mutant; Δ*hns, hns* deletion mutant (B, D) The bar plots for the expression levels of the element genes of the modulons obtained from the transcriptome data of the WT and Δ*tna* in the presence or absence of 500 μM indole (B), and the WT and Δ*hns* (D). The modulon names of the element genes are denoted below the plots. The distinct colors of the bars represent the strains and the experimental conditions as indicated in the plots. Each dot on the bars represents a single biological replicate. Statistical significance was determined by the Student's *t* test (ns, not significant; *, $p < 0.05$; **, $p < 10^{-2}$; ***, $p < 10^{-3}$; ****, $p < 10^{-4}$).

III-3-6. The modulons improve clustering the genes co-regulated regardless of the genetic backgrounds

Significantly changed activities of H-NS ($p < 10^{-2}$), flagella and chemotaxis ($p < 10^{-2}$), and putative type III secretion system (T3SS) modulons ($p < 10^{-2}$) were also observed from the transcriptome data of the WT and *hns* deletion mutant (Δ*hns*) (Fig. III-6C). The deletion of *hns* significantly decreased the activities of the H-NS modulon ($p < 10^{-2}$) (Fig. III-6C). Since *stpA* and *ecpR*, the element genes of H-NS modulon, have negative gene coefficients (Table III-3), the expressions of the genes increased significantly along with the decreased activities of the modulon in Δ*hns* (Fig. III-6D). The deletion of *hns* significantly decreased the activities of the flagella and chemotaxis modulon ($p < 10^{-2}$) (Fig. III-6C), and thereby decreased expressions of the flagella component genes *flgBCDEF* (Fig. III-6D). The deletion of *hns* also significantly increased the activities of the putative T3SS modulon ($p < 10^{-2}$) (Fig. III-6C) and thereby increased the expressions of *eivG*, the putative T3SS component gene (Fig. III-6D). The *stpA* and *ecpR*, flagella component genes, and putative T3SS component genes, known as the H-NS regulon (Lang *et al.*, 2007; Martínez-Santos *et al.*, 2012; Ueda *et al.*, 2013; Wan *et al.*, 2016), were separately classified into the H-NS, flagella and chemotaxis, and putative T3SS modulons, respectively. These results indicated that the modulons successfully clustered the inherently co-regulated genes of EHEC regardless of the genetic backgrounds.

.

III-3-7. The modulons enhance understanding on the differential expressions of the EHEC virulence and survival genes

The changed activities of the modulons were analyzed from the transcriptome data previously obtained from EHEC in the different sites of the bovine GITs in order to confirm the differential gene expressions of the pathogen in the course of infection. For example, the activities of the RpoS, flagella and chemotaxis, Stx, and LEE modulons significantly changed in the different sites of the bovine GITs (Fig. III-7A). The activities of the RpoS modulon were significantly higher in the rumen than those in other sites of the bovine GITs ($p < 10^{-2}$) (Fig. III-7A). Accordingly, the expressions of the element genes of the RpoS modulon, such as *gadABC* (Ling *et al.*, 2008), *katE* (Schellhorn, 1995), *hdeA* (Dudin *et al.*, 2013), and *slp* (Kabir *et al.*, 2004), significantly increased in the rumen (Fig. III-7B). The activities of the flagella and chemotaxis modulon, and thereby the expressions of the *flgBCDEF*, were significantly higher in the small intestine and rectum than in the rumen ($p < 0.05$) (Fig. III-7 A and B). The activities of the Stx ($p < 10^{-3}$) and LEE ($p < 0.05$) modulons, and thereby the expressions of the *stx2a*, *escE*, *escJ*, *cesL*, *sepL*, and *tir,* were significantly higher in the rectum than in other sites of the bovine GITs (Fig. III-7 A and B). Consequently, these results indicated that the activities of the modulons could successfully explain the changed expressions of the virulence and survival genes in the different sites of the bovine GITs, enhancing understanding on the spatially differentiated gene expressions of EHEC during the course of infection.

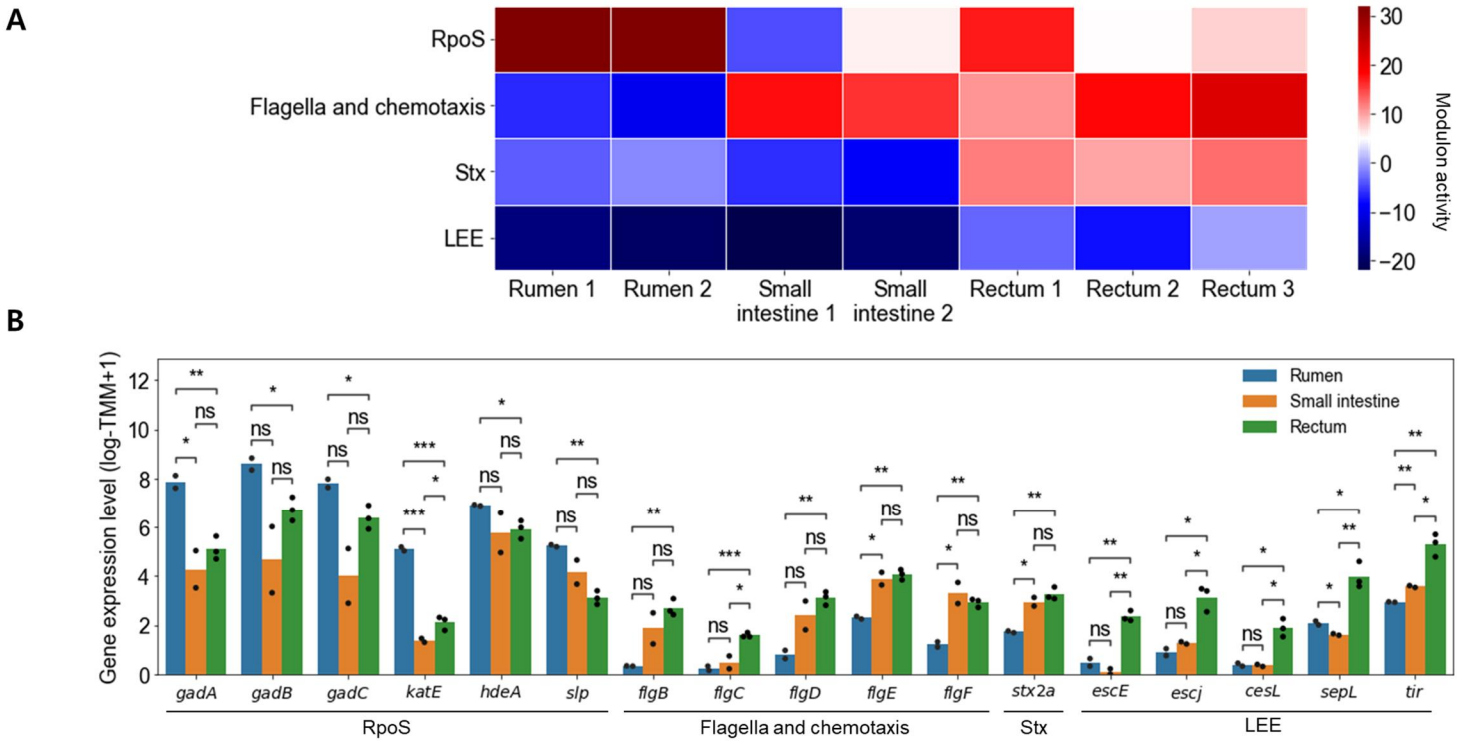

Figure III-8. The changed activities of the modulons obtained from the transcriptome data of EHEC EDL933 in the different sites of the bovine GITs. (A) Heatmap for the changed activities of the modulons in the different sites of the bovine GITs. The numbers on the bottom labels

indicate a distinct single biological replicate. Red and blue represent the high and low activity of the modulon, respectively. (B) The bar plots for the expression levels of the element genes of the modulons in the different sites of the bovine GITs. The modulon names of the element genes are denoted below the plot. The distinct colors of the bars represent the sites where EHEC was cultured as indicated in the plot. Each dot on the bars represents a single biological replicate. Statistical significance was determined by the Student's *t* test (ns, not significant; *, $p < 0.05$; **, $p < 10^{-2}$; ***, $p < 10^{-3}$).

III-4. Discussion

In this study, ICA, a machine learning method that decomposes a mixture of components into independent components, was performed to decompose the large-scale transcriptome data of EHEC into the independent sets of co-regulated genes, the modulons. As a result, the trimmed 88 transcriptome data of EHEC (Fig. III-1 A, B, C, and D) were decomposed into 64 independent modulons, which contain the target genes of the EHEC TRNs. The 64 modulons included the LEE and the Stx modulons mainly consisting of the LEE and the Stx prophage genes encoding the major virulence factors of EHEC, respectively (Fig. III-2 A and D). The activities of the LEE modulon were strongly dependent on the expression level of *ler* (Fig. III-2B), and thus the LEE modulon mostly consisted of the Ler regulon. Moreover, the LEE modulon contained additional genes such as *lpxR*, *nleA*, *stcE*, and *etpC*, which are not located in the LEE but regulated by Ler (Fig. III-2 A and C) (Grys *et al.*, 2005; Roe *et al.*, 2007; Ogawa *et al.*, 2018), indicating that ICA can precisely identify the LEE modulon to contain the target genes of the Ler TRN even not located in LEE. The Stx modulon contained the genes of the Stx prophages: CP-933V and BP-933W (Fig. III-2D). The activities of the Stx modulon were dependent on the expression levels of the antiterminator *Q*s (Fig. III-2 E and F), indicating that the Stx modulon were adequately grouped with target genes of the Stx prophage TRNs. These results suggested that ICA successfully decomposed the large-scale transcriptome data of

EHEC into the modulons.

The LEE modulon included a hypothetical Z0395 gene, which is not located within the LEE (Z5099-5141) and is not known as the Ler regulon. Interestingly, the expression of the Z0395 gene was predicted to increase along with the increased expression of *ler* (Fig. III-3A), suggesting that the Z0395 gene is a probable member of the Ler regulon. Experimentally, the deletion of *ler* significantly decreased the expression of the Z0395 gene (Fig. III-3B), confirming that the Z0395 gene in the LEE modulon is a new member of the Ler regulon. Furthermore, direct binding of Ler near the Z0395 gene was proposed by a previous ChIP-on-chip assay (32), and the Ler binding motif predicted *in silico* was found at the upstream region of the Z0395 gene (Fig. III-3 C and D) (Bailey *et al.*, 2006; Grant *et al.*, 2011). These results indicated that the Z0395 gene is a novel member of the Ler regulon, suggesting that the investigation of the modulons can discover new target genes of the current TRNs of EHEC.

The Stx modulon contained the non-prophage genes, the *thi* and *cus* locus genes, in addition to the Stx prophage genes (Fig. III-4A). The expression patterns of the *thi* and *cus* locus genes and those of other element genes in the Stx modulon were contrary (Fig. III-4 B, C, and D), and in detail, the expression levels of the *thiB*, *thiC*, and *cusC* genes have negative correlations with those of *stx2a* (Fig. III-4E). Interestingly, the levels of thiamine and copper ions known to control the expressions of the *thi* and *cus* locus genes, respectively (Vander Horn *et al.*, 1993; Webb *et al.*,

1998; Miranda-Rios *et al.*, 2001; Delmar *et al.*, 2015), regulated inversely the *stx2a* prophage gene (Fig. III-4 F and G). Considering that thiamine is mostly produced by the gut microbiota (Said *et al.*, 2001; Bhat and Kapila, 2017; Pan *et al.*, 2017), the presence of thiamine could be an environmental signal for EHEC to suppress the *thi* locus genes and to induce the Stx virulence factors in the intestinal environments. Meanwhile, copper ions which are mostly consumed with foods and then absorbed by the enterocytes in the upper small intestine are left only in trace amounts in the large intestine (45). Therefore, the relatively low copper ions also could be a signal for EHEC to suppress the *cus* locus genes and to induce the Stx virulence factors in the large intestine, the major colonization site for the pathogen (46). Consequently, the investigation of the element genes of the Stx modulon could propose novel environmental signals such as the levels of thiamine and copper ions to control expressions of the Stx prophage genes, providing further understanding of the regulation of the TRNs of EHEC virulence factors.

The TRNs of bacteria primarily consist of the genes whose expressions are regulated together by a specific growth condition or the presence of a specific TF(s) (Sastry *et al.*, 2019; DuPai *et al.*, 2020). In contrast, the modulons of bacteria consist of the genes that are identified computationally and are expressed differentially together regardless of their growth conditions and the genetic backgrounds (Saelens *et al.*, 2018; Sastry *et al.*, 2019; Tan *et al.*, 2020). Accordingly, novel gene Z0395, another element gene of the LEE modulon (Fig. III-3A), is expressed together with

the LEE genes in the presence or absence of indole (Fig. III-6 A and B). Additionally, the genes regulated by an identical TF can be classified into different modulons. For example, the flagella component genes and the putative T3SS component genes of the H-NS regulon were separately classified into the flagella and chemotaxis modulon, and putative T3SS modulon, respectively (Fig. III-6 C and D). Altogether, these results indicated that the individual modulon successfully clustered a set of genes that are inherently co-regulated under the various conditions regardless of the genetic backgrounds of EHEC.

The changed activities of the modulons can be obtained from the transcriptome data of EHEC previously observed from the different sites of the bovine GITs. The activities of the RpoS modulon including the acid resistance genes, *gadABC*, increased significantly in the rumen, the acidic environment (Fig. III-7 A and B) (Ogawa *et al.*, 2001; Ling *et al.*, 2008; Chaucheyras-Durand *et al.*, 2010). The activities of the flagella and chemotaxis modulon increased significantly in the small intestine and the rectum (Fig. III-7 A and B), which enables EHEC to move to more favorable niches (Naylor *et al.*, 2003; Xu *et al.*, 2012). The activities of the LEE and the Stx modulons increased significantly in the rectum (Fig. III-7 A and B). The LEE genes encode the crucial adherence factors for colonizing the rectum, the primary colonization site of EHEC (Naylor *et al.*, 2003). The Stxs also provide advantages for persistent colonization of EHEC by retarding the adaptive immune system at the bovine intestinal mucosa (Menge, 2020). Altogether, these results indicated that the

changed activities of the modulons obtained from the transcriptome data could successfully explain the pathogenesis of EHEC during the course of infection in bovine.

In summary, ICA of the large-scale transcriptome data identified the modulons consisting of the target genes of the EHEC TRNs. Further analysis of the modulons revealed that the Z0395 gene and the *thi* and *cus* locus genes are novel element genes of the LEE and Stx modulons, respectively. Concurrently, the Stx prophage genes were also regulated by thiamine and copper ions controlling the *thi* and *cus* locus genes, respectively. Changed activities of the modulons consisting of the inherently co-regulated genes enhanced understanding on the differential expressions of the EHEC virulence and survival genes in response to specific intestinal environments. Consequently, ICA can expand and enhance the current understating of the TRNs of EHEC, suggesting that ICA can provide broader insight into the TRNs of other pathogens from their transcriptome data.

Chapter IV.

Conclusion

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 is getting attention because of the recent advances in ML algorithms and sequencing technologies. Here, I 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 repertories 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.

The elucidation of the transcriptional regulatory networks (TRNs) of enterohemorrhagic *Escherichia coli* (EHEC) is critical to understanding its pathogenesis and survival in the host. However, the analyses of current TRNs are still limited to comprehensively understanding their target genes generally co-regulated under various conditions regardless of the genetic backgrounds. In this study, independent component analysis (ICA), a machine learning-based decomposition method, was used to decompose the large-scale transcriptome data of EHEC into the modulons, which contain the target genes of several TRNs. The locus of enterocyte effacement (LEE) and the Shiga toxin (Stx) modulons mainly consisted of the Ler regulon and the Stx prophage genes, respectively, confirming that ICA properly grouped the co-regulated genes of EHEC. Further investigation revealed that the LEE modulon contained the Z0395 gene as a novel member of the Ler regulon, and the Stx modulon contained the *thi* and *cus* locus genes in addition to the Stx prophage genes. Concurrently, the Stx prophage genes were also regulated by thiamine and copper ions known to control the *thi* and *cus* locus genes, respectively. The modulons effectively clustered the genes co-regulated regardless of the growth conditions and the genetic backgrounds of EHEC. The changed activities of the individual modulons successfully explained the differential expressions of the virulence and survival genes during the course of infection in bovine. Altogether, these results suggested that ICA of the large-scale transcriptome data can expand and enhance the current understanding of the TRNs of EHEC.

References

Abadi, M., Barham, P., and Chen, J. (2016). TensorFlow: A System for Large-Scale Machine Learning. *12th {USENIX} Symp. Oper. Syst. Des. Implement. ({OSDI} 16)*.

Abe, H., Miyahara, A., Oshima, T., Tashiro, K., Ogura, Y., Kuhara, S., et al. (2008). Global regulation by horizontally transferred regulators establishes the pathogenicity of Escherichia coli. *DNA Res.* 15, 13–23. doi:10.1093/dnares/dsm028.

Abu-Ali, G. S., Lacher, D. W., Wick, L. M., Qi, W., and Whittam, T. S. (2009). Genomic diversity of pathogenic Escherichia coli of the EHEC 2 clonal complex. *BMC Genomics* 10, 1–16. doi:10.1186/1471-2164-10-296.

Adams, N., Byrne, L., Edge, J., Hoban, A., Jenkins, C., and Larkin, L. (2019). Gastrointestinal infections caused by consumption of raw drinking milk in England & Wales, 1992–2017. *Epidemiol. Infect.* 147, e281. doi:10.1017/S095026881900164X.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 1340–1347. doi:10.1093/bioinformatics/btq134.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21. doi:10.1093/nar/gkw387.

Atlung, T., and Ingmer, H. (1997). H-NS: a modulator of environmentally regulated gene expression. *Mol. Microbiol.* 24, 7–17. doi:10.1046/j.1365-2958.1997.3151679.x.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi:10.1093/nar/gkl198.

Baum, D. A., Smith, S. D., and Donovan, S. S. S. (2005). The Tree-Thinking

Challenge. *Science (80-.).* 310, 979–980. doi:10.1126/science.1117727.

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., and Feil, E. J. (2019). PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience* 8, 1–9. doi:10.1093/gigascience/giz119.

Belgiu, M., and Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* 114, 24–31. doi:10.1016/j.isprsjprs.2016.01.011.

Berrar, D., and Flach, P. (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Brief. Bioinform.* 13, 83–97. doi:10.1093/bib/bbr008.

Bhat, M. I., and Kapila, R. (2017). Dietary metabolites derived from gut microbiota: critical modulators of epigenetic changes in mammals. *Nutr. Rev.* 75, 374–389. doi:10.1093/nutrit/nux001.

Bhavsar, H., and Panchal, M. H. (2012). A Review on Support Vector Machine for Data Classification. *Int. J. Adv. Res. Comput. Eng. Technol.* 1, 2278–1323.

Boisen, N., Melton-Celsa, A. R., Scheutz, F., O'Brien, A. D., and Nataro, J. P. (2015). Shiga toxin 2a and Enteroaggregative Escherichia coli – a deadly combination. *Gut Microbes* 6, 272–278. doi:10.1080/19490976.2015.1054591.

Breiman, L. (2020). Random Forests. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 12343 LNCS, 503–515. doi:doi.org/10.1023/A:1010933404324.

Bremer, E., and Krämer, R. (2019). Responses of Microorganisms to Osmotic Stress. *Annu. Rev. Microbiol.* 73, 313–334. doi:10.1146/annurev-micro-020518-115504.

Brüssow, H., Canchaya, C., and Hardt, W.-D. (2004). Brüssow, Canchaya, Hardt - 2004 - Phages and the Evolution of Bacterial Pathogens from Genomic Rearrangements to Lysogenic Conversion.pdf. doi:10.1128/MMBR.68.3.560-602.2004.

Brynildsrud, O., Bohlin, J., Scheffer, L., and Eldholm, V. (2016). Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* 17, 238. doi:10.1186/s13059-016-1108-8.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment

using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nat. Methods* 15, 233–234. doi:10.1038/nmeth.4642.

Caprioli, A., Morabito, S., Brugère, H., and Oswald, E. (2005). Enterohaemorrhagic Escherichia coli : emerging issues on virulence and modes of transmission. *Vet. Res.* 36, 289–311. doi:10.1051/vetres:2005002.

Casjens, S. R., and Hendrix, R. W. (2015). Bacteriophage lambda: Early pioneer and still relevant. *Virology* 479–480, 310–330. doi:10.1016/j.virol.2015.02.010.

Chaucheyras-Durand, F., Faqir, F., Ameilbonne, A., Rozand, C., and Martin, C. (2010). Fates of Acid-Resistant and Non-Acid-Resistant Shiga Toxin-Producing Escherichia coli Strains in Ruminant Digestive Contents in the Absence and Presence of Probiotics. *Appl. Environ. Microbiol.* 76, 640–647. doi:10.1128/AEM.02054-09.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6. doi:10.1186/s12864-019-6413-7.

Collobert, R., and Bengio, S. (2004). Links between perceptrons, MLPs and SVMs. in *Twenty-first international conference on Machine learning - ICML '04* (New York, New York, USA: ACM Press), 23. doi:10.1145/1015330.1015415.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi:10.1007/BF00994018.

D'agostino, R. B., Belanger, A., and D'agostino, R. B. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *Am. Stat.* 44, 316–321. doi:10.1080/00031305.1990.10475751.

Delmar, J. A., Su, C. C., and Yu, E. W. (2015). Heavy metal transport by the CusCFBA efflux system. *Protein Sci.* 24, 1720–1736. doi:10.1002/pro.2764.

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi:10.1016/j.jneumeth.2003.10.009.

Dobin, A., Davis, C. A. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Dudin, O., Lacour, S., and Geiselmann, J. (2013). Expression dynamics of RpoS/Crl-dependent genes in Escherichia coli. *Res. Microbiol.* 164, 838–847. doi:10.1016/j.resmic.2013.07.002.

DuPai, C. D., Wilke, C. O., and Davies, B. W. (2020). A Comprehensive Coexpression Network Analysis in Vibrio cholerae. *mSystems* 5, 1–12. doi:10.1128/mSystems.00550-20.

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., et al. (2005). Diversity of the Human Intestinal Microbial Flora. *Science (80-.).* 308, 1635–1638. doi:10.1126/science.1110591.

Eichhorn, I., Heidemanns, K., Semmler, T., Kinnemann, B., Mellmann, A., Harmsen, D., et al. (2015). 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. doi:10.1128/AEM.01921-15.

Eisenstein, B. I., and Dodd, D. C. (1982). Pseudocatabolite repression of type 1 fimbriae of Escherichia coli. *J. Bacteriol.* 151, 1560–1567. doi:10.1128/jb.151.3.1560-1567.1982.

Elliott, S. J., Sperandio, V., Giron, J. A., Shin, S., Mellies, J. L., Wainwright, L., et al. (2000). The locus of enterocyte effacement (LEE)-encoded regulator controls expression of both LEE- and non-LEE-encoded virulence factors in enteropathogenic and enterohemorrhagic Escherichia coli. *Infect. Immun.* 68, 6115–6126. doi:10.1128/IAI.68.11.6115-6126.2000.

European Food Safety Authority (2013). Scientific Opinion on VTEC-seropathotype and scientific criteria regarding pathogenicity assessment. *EFSA J.* 11, 3138. doi:10.2903/j.efsa.2013.3138.

Fang, X., Sastry, A., Mih, N., Kim, D., Tan, J., Yurkovich, J. T., et al. (2017). Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proc. Natl. Acad. Sci.* 114, 10286–10291. doi:10.1073/pnas.1702581114.

Franz, E., Delaquis, P., Morabito, S., Beutin, L., Gobius, K., Rasko, D. A., et al. (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.* 187, 57–72. doi:10.1016/j.ijfoodmicro.2014.07.002.

Fratamico, P. M., DebRoy, C., Liu, Y., Needleman, D. S., Baranzoni, G. M., and Feng, P. (2016). Advances in Molecular Serotyping and Subtyping of Escherichia coli†. *Front. Microbiol.* 7, 1–8. doi:10.3389/fmicb.2016.00644.

Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., et al. (2011). RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.* 39, D98–D105. doi:10.1093/nar/gkq1110.

Gao, Y.-D., Zhao, Y., and Huang, J. (2014). Metabolic Modeling of Common Escherichia coli Strains in Human Gut Microbiome. *Biomed Res. Int.* 2014, 1–11. doi:10.1155/2014/694967.

Gould, L. H., Mody, R. K., Ong, K. L., Clogher, P., Cronquist, A. B., Garman, K. N., et al. (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. doi:10.1089/fpd.2012.1401.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi:10.1093/bioinformatics/btr064.

Grys, T. E., Siegel, M. B., Lathem, W. W., and Welch, R. A. (2005). The StcE Protease Contributes to Intimate Adherence of Enterohemorrhagic Escherichia coli O157:H7 to Host Cells. *Infect. Immun.* 73, 1295–1303. doi:10.1128/IAI.73.3.1295-1303.2005.

Guoshen Yu, Sapiro, G., and Mallat, S. (2012). Solving Inverse Problems With Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity. *IEEE Trans. Image Process.* 21, 2481–2499. doi:10.1109/TIP.2011.2176743.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi:10.1093/bioinformatics/btt086.

Hald, T., Aspinall, W., Devleesschauwer, B., Cooke, R., Corrigan, T., Havelaar, A. H., et al. (2016). World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation. *PLoS One* 11, e0145839. doi:10.1371/journal.pone.0145839.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi:10.1038/s41586-020-2649-2.

Havelaar, A. H., Kirk, M. D., Torgerson, P. R., Gibb, H. J., Hald, T., Lake, R. J., et al. (2015). World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. *PLOS Med.* 12, e1001923. doi:10.1371/journal.pmed.1001923.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: The next challenge. *Nat. Rev. Genet.* 11, 855–866. doi:10.1038/nrg2897.

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi:10.1093/molbev/msx148.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks* 13, 411–430. doi:10.1016/S0893-6080(00)00026-5.

Jahromi, A. H., and Taheri, M. (2018). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. *2017 Artif. Intell. Signal Process. Conf.* 2018-Janua, 209–212. doi:10.1109/AISP.2017.8324083.

James, C. J., and Hesse, C. W. (2005). Independent component analysis for biomedical signals. *Physiol. Meas.* 26, 15–39. doi:10.1088/0967-3334/26/1/R02.

Jang, K. K., Lee, Z.-W., Kim, B., Jung, Y. H., Han, H. J., Kim, M. H., et al. (2017). Identification and characterization of Vibrio vulnificus plpA encoding a phospholipase A2 essential for pathogenesis. *J. Biol. Chem.* 292, 17129–17143.

doi:10.1074/jbc.M117.791657.

Jiang, Y., Chen, B., Duan, C., Sun, B., Yang, J., and Yang, S. (2015). Multigene editing in the Escherichia coli genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* 81, 2506–2514. doi:10.1128/AEM.04023-14.

Joensen, K. G., Tetzschner, A. M. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data. *J. Clin. Microbiol.* 53, 2410–2426. doi:10.1128/JCM.00008-15.

Johannes, L., and Römer, W. (2010). Shiga toxins — from cell biology to biomedical applications. *Nat. Rev. Microbiol.* 8, 105–116. doi:10.1038/nrmicro2279.

Johansen, B. K., Wasteson, Y., Granum, P. E., and Brynestad, S. (2001). Mosaic structure of Shiga-toxin-2-encoding phages isolated from Escherichia coli O157:H7 indicates frequent gene exchange between lambdoid phage genomes. *Microbiology* 147, 1929–1936. doi:10.1099/00221287-147-7-1929.

John, G. H., and Langley, P. (2013). Estimating Continuous Distributions in Bayesian Classifiers. *Proc. Elev. Conf. Uncertain. Artif. Intell.*, 338–345. Available at: http://arxiv.org/abs/1302.4964.

Jønsson, R., Struve, C., Boll, E. J., Boisen, N., Joensen, K. G., Sørensen, C. A., et al. (2017). A novel pAA virulence plasmid encoding toxins and two distinct variants of the fimbriae of enteroaggregative Escherichia coli. *Front. Microbiol.* 8. doi:10.3389/fmicb.2017.00263.

Kabir, M. S., Sagara, T., Oshima, T., Kawagoe, Y., Mori, H., Tsunedomi, R., et al. (2004). Effects of mutations in the rpoS gene on cell viability and global gene expression under nitrogen starvation in Escherichia coli. *Microbiology* 150, 2543–2553. doi:10.1099/mic.0.27012-0.

Kalivoda, K. A., Steenbergen, S. M., and Vimr, E. R. (2013). Control of the Escherichia coli Sialoregulon by Transcriptional Repressor NanR. *J. Bacteriol.* 195, 4689–4701. doi:10.1128/JB.00692-13.

Kalivoda, K. A., Steenbergen, S. M., Vimr, E. R., and Plumbridge, J. (2003). Regulation of Sialic Acid Catabolism by the DNA Binding Protein NanR in Escherichia coli. *J. Bacteriol.* 185, 4806–4815. doi:10.1128/JB.185.16.4806-4815.2003.

Kaper, J. B., Nataro, J. P., and Mobley, H. L. T. (2004). Pathogenic Escherichia coli. *Nat. Rev. Microbiol.* 2, 123–140. doi:10.1038/nrmicro818.

Karmali, M. A. (2017). 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. doi:10.1093/cid/ciw708.

Kelly, M., Hart, E., Mundy, R., Marchès, O., Wiles, S., Badea, L., et al. (2006). Essential role of the type III secretion system effector NleB in colonization of mice by Citrobacter rodentium. *Infect. Immun.* 74, 2328–2337. doi:10.1128/IAI.74.4.2328-2337.2006.

Kenny, B., DeVinney, R., Stein, M., Reinscheid, D. J., Frey, E. A., and Finlay, B. B. (1997). Enteropathogenic E. coli (EPEC) transfers its receptor for intimate adherence into mammalian cells. *Cell* 91, 511–520. doi:10.1016/S0092-8674(00)80437-7.

Kijewski, A., Witsø, I. L., Iversen, H., Rønning, H. T., L'Abée-Lund, T., Wasteson, Y., et al. (2020). Vitamin K Analogs Influence the Growth and Virulence Potential of Enterohemorrhagic Escherichia coli. *Appl. Environ. Microbiol.* 86, 1–16. doi:10.1128/AEM.00583-20.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artif. Intell. Rev.* 39, 261–283. doi:10.1007/s10462-011-9272-4.

Koutsoumanis, K., Allende, A., Alvarez-Ordóñez, A., Bover-Cid, S., Chemaly, M., Davies, R., et al. (2020). 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. doi:10.2903/j.efsa.2020.5967.

Koza, J. R., Bennett, F. H., Andre, D., and Keane, M. A. (1996). "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming," in *Artificial Intelligence in Design '96* (Dordrecht: Springer Netherlands), 151–170. doi:10.1007/978-94-009-0279-4_9.

Kumar, A., and Sperandio, V. (2019). Indole Signaling at the Host-Microbiota-Pathogen Interface. *MBio* 10. doi:10.1128/mBio.01031-19.

Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C. O., et al. (2007). High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Res.* 35,

6330–6337. doi:10.1093/nar/gkm712.

Lehnherr, H., Maguin, E., Jafri, S., and Yarmolinsky, M. B. (1993). 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. doi:10.1006/jmbi.1993.1521.

Levine, M. M. (1987). Escherichia coli that Cause Diarrhea: Enterotoxigenic, Enteropathogenic, Enteroinvasive, Enterohemorrhagic, and Enteroadherent. *J. Infect. Dis.* 155, 377–389. doi:10.1093/infdis/155.3.377.

Li, M., Rosenshine, I., Yu, H. B., Nadler, C., Mills, E., Hew, C. L., et al. (2006). Identification and characterization of NleI, a new non-LEE-encoded effector of enteropathogenic Escherichia coli (EPEC). *Microbes Infect.* 8, 2890–2898. doi:10.1016/j.micinf.2006.09.006.

Lim, J. Y., Yoon, J. W., and Hovde, C. J. (2010). A brief overview of Escherichia coli O157:H7 and its plasmid O157. *J. Microbiol. Biotechnol.* 20, 1–10. doi:10.4014/jmb.0908.08007.

Ling, J., Sharma, M., and Bhagwat, A. A. (2008). Role of RNA polymerase sigma-factor (RpoS) in induction of glutamate-dependent acid-resistance of Escherichia albertii under anaerobic conditions. *FEMS Microbiol. Lett.* 283, 75–82. doi:10.1111/j.1574-6968.2008.01153.x.

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 47, D687–D692. doi:10.1093/nar/gky1080.

Loh, W. (2011). Classification and regression trees. *WIREs Data Min. Knowl. Discov.* 1, 14–23. doi:10.1002/widm.8.

Lupolova, N., Dallman, T. J., Holden, N. J., and Gally, D. L. (2017). Patchy promiscuity: Machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. *Microb. Genomics* 3, 1–10. doi:10.1099/mgen.0.000135.

Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L., and Gally, D. L. (2016). 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. doi:10.1073/pnas.1606567113.

Lupolova, N., Lycett, S. J., and Gally, D. L. (2019). A guide to machine learning for bacterial host attribution using genome sequence data. *Microb. Genomics* 5. doi:10.1099/mgen.0.000317.

Martínez-Santos, V. I., Medrano-López, A., Saldaña, Z., Girón, J. A., and Puente, J. L. (2012). Transcriptional Regulation of the ecp Operon by EcpR, IHF, and H-NS in Attaching and Effacing Escherichia coli. *J. Bacteriol.* 194, 5020–5033. doi:10.1128/JB.00915-12.

Martins, F. H., Kumar, A., Abe, C. M., Carvalho, E., Nishiyama-Jr, M., Xing, C., et al. (2020). EspFu-Mediated Actin Assembly Enhances Enteropathogenic Escherichia coli Adherence and Activates Host Cell Inflammatory Signaling Pathways. *MBio* 11, 1–18. doi:10.1128/mBio.00617-20.

Martins, F. H., Nepomuceno, R., Piazza, R. M. F., and Elias, W. P. (2017). Phylogenetic distribution of tir-cytoskeleton coupling protein (tccP and tccP2) genes in atypical enteropathogenic Escherichia coli. *FEMS Microbiol. Lett.* 364, 1–7. doi:10.1093/femsle/fnx101.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference*, 56–61. doi:10.25080/Majora-92bf1922-00a.

Mellies, J. L., Elliott, S. J., Sperandio, V., Donnenberg, M. S., and Kaper, J. B. (1999). The Per regulon of enteropathogenic Escherichia coli: Identification of a regulatory cascade and a novel transcriptional activator, the locus of enterocyte effacement (LEE)-encoded regulator (Ler). *Mol. Microbiol.* 33, 296–306. doi:10.1046/j.1365-2958.1999.01473.x.

Menge, C. (2020). The Role of Escherichia coli Shiga Toxins in STEC Colonization of Cattle. *Toxins (Basel).* 12, 607. doi:10.3390/toxins12090607.

Miranda-Rios, J., Navarro, M., and Soberon, M. (2001). A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci.* 98, 9736–9741. doi:10.1073/pnas.161168098.

Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLoS Comput. Biol.* 14, 1–17. doi:10.1371/journal.pcbi.1006258.

Morin, N., Tirling, C., Ivison, S. M., Kaur, A. P., Nataro, J. P., and Steiner, T. S. (2010). Autoactivation of the AggR regulator of enteroaggregative Escherichia coli in vitro and in vivo. *FEMS Immunol. Med. Microbiol.* 58, 344–355. doi:10.1111/j.1574-695X.2009.00645.x.

Naseer, U., Løbersli, I., Hindrum, M., Bruvik, T., and Brandal, L. T. (2017). 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. doi:10.1007/s10096-017-2974-z.

Naylor, S. W., Low, J. C., Besser, T. E., Mahajan, A., Gunn, G. J., Pearce, M. C., et al. (2003). Lymphoid Follicle-Dense Mucosa at the Terminal Rectum Is the Principal Site of Colonization of Enterohemorrhagic Escherichia coli O157:H7 in the Bovine Host. *Infect. Immun.* 71, 1505–1512. doi:10.1128/IAI.71.3.1505-1512.2003.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi:10.1093/nar/gkv1189.

Ogawa, M., Shimizu, K., Nomoto, K., Tanaka, R., Hamabata, T., Yamasaki, S., et al. (2001). Inhibition of in vitro growth of Shiga toxin-producing Escherichia coli O157:H7 by probiotic Lactobacillus strains due to production of lactic acid. *Int. J. Food Microbiol.* 68, 135–140. doi:10.1016/S0168-1605(01)00465-2.

Ogawa, R., Yen, H., Kawasaki, K., and Tobe, T. (2018). Activation of lpxR gene through enterohaemorrhagic Escherichia coli virulence regulators mediates lipid A modification to attenuate innate immune response. *Cell. Microbiol.* 20, 1–13. doi:10.1111/cmi.12806.

Ogura, Y., Ooka, T., Asadulghani, Terajima, J., Nougayrède, J.-P., Kurokawa, K., et al. (2007). Extensive genomic diversity and selective conservation of virulence-determinants in enterohemorrhagic Escherichia coli strains of O157 and non-O157 serotypes. *Genome Biol.* 8, R138. doi:10.1186/gb-2007-8-7-r138.

Okhuysen, P. C., and DuPont, H. L. (2010). Enteroaggregative Escherichia coli (EAEC): A Cause of Acute and Persistent Diarrhea of Worldwide Importance.

J. Infect. Dis. 202, 503–505. doi:10.1086/654895.

Otsuka, Y., and Yonesaki, T. (2012). Dmd of bacteriophage T4 functions as an antitoxin against Escherichia coli LsoA and RnlA toxins. *Mol. Microbiol.* 83, 669–681. doi:10.1111/j.1365-2958.2012.07975.x.

Pacheco, A. R., and Sperandio, V. (2012). Shiga toxin in enterohemorrhagic E.coli: regulation and novel anti-virulence strategies. *Front. Cell. Infect. Microbiol.* 2, 97–109. doi:10.3389/fcimb.2012.00081.

Pan, X., Xue, F., Nan, X., Tang, Z., Wang, K., Beckers, Y., et al. (2017). Illumina Sequencing Approach to Characterize Thiamine Metabolism Related Bacteria and the Impacts of Thiamine Supplementation on Ruminal Microbiota in Dairy Cows Fed High-Grain Diets. *Front. Microbiol.* 8, 1–10. doi:10.3389/fmicb.2017.01818.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* 2, 559–572. doi:10.1080/14786440109462720.

Pieper, R., Zhang, Q., Clark, D. J., Parmar, P. P., Alami, H., Suh, M.-J., et al. (2013). Proteomic View of Interactions of Shiga Toxin-Producing Escherichia coli with the Intestinal Environment in Gnotobiotic Piglets. *PLoS One* 8, e66462. doi:10.1371/journal.pone.0066462.

Platenkamp, A., and Mellies, J. L. (2018). Environment Controls LEE Regulation in Enteropathogenic Escherichia coli. *Front. Microbiol.* 9, 1–15. doi:10.3389/fmicb.2018.01694.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. large margin Classif.* 10, 61–74.

Prieto, A., Bernabeu, M., Sánchez-Herrero, J. F., Pérez-Bosque, A., Miró, L., Bäuerl, C., et al. (2021). Modulation of AggR levels reveals features of virulence regulation in enteroaggregative E. coli. *Commun. Biol.* 4, 1295. doi:10.1038/s42003-021-02820-9.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for

differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25. doi:10.1186/gb-2010-11-3-r25.

Roe, A. J., Tysall, L., Dransfield, T., Wang, D., Fraser-Pitt, D., Mahajan, A., et al. (2007). Analysis of the expression, regulation and export of NleA-E in Escherichia coli O157:H7. *Microbiology* 153, 1350–1360. doi:10.1099/mic.0.2006/003707-0.

Römer, W., Berland, L., Chambon, V., Gaus, K., Windschiegl, B., Tenza, D., et al. (2007). Shiga toxin induces tubular membrane invaginations for its uptake into cells. *Nature* 450, 670–675. doi:10.1038/nature05996.

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9, 1090. doi:10.1038/s41467-018-03424-4.

Said, H. M., Ortiz, A., Subramanian, V. S., Neufeld, E. J., Moyer, M. P., and Dudeja, P. K. (2001). Mechanism of thiamine uptake by human colonocytes: studies with cultured colonic epithelial cell line NCM460. *Am. J. Physiol. Liver Physiol.* 281, G144–G150. doi:10.1152/ajpgi.2001.281.1.G144.

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10, 1–14. doi:10.1038/s41467-019-13483-w.

Schellhorn, H. E. (1995). Regulation of hydroperoxidase (catalase) expression in Escherichia coli. *FEMS Microbiol. Lett.* 131, 113–119. doi:10.1111/j.1574-6968.1995.tb07764.x.

Scheutz, F., Teel, L. D., Beutin, L., Pierard, D., Buvens, G., Karch, H., et al. (2012). Multicenter Evaluation of a Sequence-Based Protocol for Subtyping Shiga Toxins and Standardizing Stx Nomenclature. *J. Clin. Microbiol.* 50, 2951–2963. doi:10.1128/JCM.00860-12.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153.

Sharan, S. K., Thomason, L. C., Kuznetsov, S. G., and Court, D. L. (2009). Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.* 4, 206–223. doi:10.1038/nprot.2008.227.

Sheikh, J., Czeczulin, J. R., Harrington, S., Hicks, S., Henderson, I. R., Le

Bouguénec, C., et al. (2002). A novel dispersin protein in enteroaggregative Escherichia coli. *J. Clin. Invest.* 110, 1329–1337. doi:10.1172/JCI16172.

Sheng, H., Lim, J. Y., Knecht, H. J., Li, J., and Hovde, C. J. (2006). Role of Escherichia coli O157:H7 Virulence Factors in Colonization at the Bovine Terminal Rectal Mucosa. *Infect. Immun.* 74, 4685–4693. doi:10.1128/IAI.00406-06.

Smith, J. L., and Fratamico, P. M. (2018). Emerging and Re-Emerging Foodborne Pathogens. *Foodborne Pathog. Dis.* 15, 737–757. doi:10.1089/fpd.2018.2493.

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

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi:10.1093/bioinformatics/btu033.

Steyert, S. R., Sahl, J. W., Fraser, C. M., Teel, L. D., Scheutz, F., and Rasko, D. A. (2012). Comparative Genomics and stx Phage Characterization of LEE-Negative Shiga Toxin-Producing Escherichia coli. *Front. Cell. Infect. Microbiol.* 2, 133. doi:10.3389/fcimb.2012.00133.

Svensson, C. M., Krusekopf, S., Lücke, J., and Thilo Figge, M. (2014). Automated detection of circulating tumor cells with naive Bayesian classifiers. *Cytom. Part A* 85, 501–511. doi:10.1002/cyto.a.22471.

Sy, B. M., Lan, R., and Tree, J. J. (2020). Early termination of the Shiga toxin transcript generates a regulatory small RNA. *Proc. Natl. Acad. Sci. U. S. A.* 117, 25055–25065. doi:10.1073/pnas.2006730117.

Tan, J., Sastry, A. V., Fremming, K. S., Bjørn, S. P., Hoffmeyer, A., Seo, S., et al. (2020). Independent component analysis of E. coli's transcriptome reveals the cellular processes that respond to heterologous gene expression. *Metab. Eng.* 61, 360–368. doi:10.1016/j.ymben.2020.07.002.

Tauxe, R. V. (2002). Emerging foodborne pathogens. *Int. J. Food Microbiol.* 78, 31–41. doi:10.1016/S0168-1605(02)00232-5.

Tobe, T., Beatson, S. A., Taniguchi, H., Abe, H., Bailey, C. M., Fivian, A., et al. (2006). An extensive repetoire of type III secretion effectors in Escherichia coli

O157 and the role of lambdoid phages in their dissemination. *Proc. Natl. Acad. Sci. U. S. A.* 103, 14941–14946. doi:10.1073/pnas.0604891103.

Toloşi, L., and Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* 27, 1986–1994. doi:10.1093/bioinformatics/btr300.

Torres, A. G., López-Sánchez, G. N., Milflores-Flores, L., Patel, S. D., Rojas-López, M., Martínez de la Peña, C. F., et al. (2007). Ler and H-NS, Regulators Controlling Expression of the Long Polar Fimbriae of Escherichia coli O157:H7. *J. Bacteriol.* 189, 5916–5928. doi:10.1128/JB.00245-07.

Ueda, T., Takahashi, H., Uyar, E., Ishikawa, S., Ogasawara, N., and Oshima, T. (2013). Functions of the Hha and YdgT Proteins in Transcriptional Silencing by the Nucleoid Proteins, H-NS and StpA, in Escherichia coli. *DNA Res.* 20, 263–271. doi:10.1093/dnares/dst008.

UniProt Consortium, and Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049.

Vander Horn, P. B., Backstrom, A. D., Stewart, V., and Begley, T. P. (1993). Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in Escherichia coli K-12. *J. Bacteriol.* 175, 982–992. doi:10.1128/jb.175.4.982-992.1993.

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., and Mueller, A. (2015). Scikit-learn. *GetMobile Mob. Comput. Commun.* 19, 29–33. doi:10.1145/2786984.2786995.

Vimr, E. R. (2013). Unified Theory of Bacterial Sialometabolism: How and Why Bacteria Metabolize Host Sialic Acids. *ISRN Microbiol.* 2013, 1–26. doi:10.1155/2013/816713.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2.

Vouga, M., and Greub, G. (2016). Emerging bacterial pathogens: the past and beyond. *Clin. Microbiol. Infect.* 22, 12–21. doi:10.1016/j.cmi.2015.10.010.

Wan, B., Zhang, Q., Tao, J., Zhou, A., Yao, Y., and Ni, J. (2016). Global transcriptional regulation by H-NS and its biological in fl uence on the virulence of Enterohemorrhagic Escherichia coli. *Gene* 588, 115–123.

doi:10.1016/j.gene.2016.05.007.

Webb, E., Claas, K., and Downs, D. (1998). thiBPQ Encodes an ABC Transporter Required for Transport of Thiamine and Thiamine Pyrophosphate in Salmonella typhimurium. *J. Biol. Chem.* 273, 8946–8950. doi:10.1074/jbc.273.15.8946.

Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 10, 618–630. doi:10.1038/nrmicro2852.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. doi:10.1186/s13059-019-1891-0.

World Health Organization (2011). Enterohaemorrhagic Escherichia coli in raw beef and beef products: approaches for the provision of scientific advice. World Health Organization.

World Health Organization (2018). Shiga toxin-producing Escherichia coli (STEC) and food: attribution, characterization, and monitoring: report. World Health Organization.

World Health Organization, Risk, M., and Series, A. (2018). *Shiga toxin-producing Escherichia coli (STEC) and food: attribution, characterization, and monitoring: report*. World Health Organization.

Xu, X., McAteer, S. P., Tree, J. J., Shaw, D. J., Wolfson, E. B. K., Beatson, S. A., et al. (2012). Lysogeny with Shiga Toxin 2-Encoding Bacteriophages Represses Type III Secretion in Enterohemorrhagic Escherichia coli. *PLoS Pathog.* 8, e1002672. doi:10.1371/journal.ppat.1002672.

Yang, X.-S., Lee, S., Lee, S., and Theera-Umpon, N. (2015). Information Analysis of High-Dimensional Data and Applications. *Math. Probl. Eng.* 2015, 1–2. doi:10.1155/2015/126740.

Young-sun et al., Y. (2019). 국내 수인성,식품매개감염병 병원체감시에 따른 병원성대장균 현황, 2010~2019. 질병관리청.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. Y. (2017). Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods Ecol. Evol.* 8, 28–36. doi:10.1111/2041-210X.12628.

Zatyka, M., and Thomas, C. M. (1998). Control of genes for conjugative transfer of

plasmids and other mobile elements. *FEMS Microbiol. Rev.* 21, 291–319. doi:10.1111/j.1574-6976.1998.tb00355.x.

국문초록

장출혈성대장균(enterohemorrhagic *Escherichia coli*, EHEC)은 묽은 설사에서부터 신장에 영구적인 손상에 이르기까지 다양한 질병을 일으키는 병원균이다. 장출혈성대장균은 새로운 균주가 지속적으로 출현해 전세계적으로 식중독 사태를 일으키고 있어 주요한 공중보건문제를 야기하는 식중독균으로 여겨진다. 이처럼 새로이 출현하는 장출혈성대장균을 예방하고 관리하기 위해서는 균이 가지는 잠재적인 병원성을 정확하게 평가할 필요가 있다. 그러나, 잠재적 병원성을 평가하기 위해 사용되는 기존의 혈청형 및 독성유전자조합 방법은 새로운 혈청형 또는 독성유전자조합을 가지는 신종 장출혈성대장균의 병원성을 평가하는 데 한계가 있다. 본 연구는 기계학습(machine learning) 기술을 활용하여 장출혈성대장균이 가지는 전장유전체 정보(whole genome sequencing data)를 활용해 균주의 잠재적 병원성을 평가할 수 있는 support vector machine (SVM) 모델을 개발했다. 개발한 SVM 모델은 기존에 장출혈성대장균 감염을 많이 일으킨 분리원으로부터 분리된 균주들의 병원성을 성공적으로 예측했을 뿐만 아니라, 발병 이력이 있는 균주 및 기존의 방법으로 평가할 수 없는 균주들의 병원성 또한 정확하게 예측했다. 또한, SVM 모델은 잠재적 병원성의 유무만을 예측하는 기존 방법에서 더 나아가, 개별 균주들이 가지는 병원성 정도의 차이를 보다 세밀하게 구분할 수 있었다. 순열중요도분석(permutation importance analysis)을 통해 SVM 모델을 분석한 결과, 병원성에 기여할 것이라 예측되는 중요 유전자 후보들을 발굴할 수 있었다. 결과적으로, 본 연구는 장출혈성대장균의 잠재적 병원성을 예측할 수 있는 새로운 방법인 SVM 모델을 제시했으며, 이를

통해 병원성에 기여할 수 있는 후보 유전자들을 발굴했다. 한편, 다양한 장출혈성대장균에서 널리 보존된 전사조절 네트워크(transcriptional regulatory network, TRN)를 파악하는 것은 새로이 출현하는 장출혈성대장균에의한 감염을 예방하고 치료하기 위해 필요하다. 그러나, 현재까지의 장출혈성대장균 전사조절 네트워크 연구들은 균주의 유전적 배경과 상관없이 다양한 환경에서 공통적으로 조절되는 유전자들을 포괄적으로 동정하고 분석하는데 어려움이 있다. 본 연구에서는 독립성분분석(independent component analysis, ICA) 알고리즘을 사용하는 기계학습 모델인 ICA 모델을 개발했으며, 이를 장출혈성대장균의 대규모 전사체 정보(transcriptome data)에 적용해 각각 독립적으로 공동 조절되는 유전자들의 집합인 모듈론(modulon)을 식별했다. 모듈론 중에는 장출혈성대장균의 주요 독성인자를 암호화하는 LEE (locus of enterocyte effacement) 유전자들로 주로 구성된 LEE 모듈론과 시가독소(Shiga toxin, Stx)를 포함하는 프로파지(prophage) 유전자들로 주로 구성된 Stx 모듈론이 포함돼 있었다. 이는 ICA 모델을 통해 공동 조절되는 유전자 집합을 적절히 식별할 수 있음을 의미한다. LEE 모듈론을 추가 분석한 결과, LEE 모듈론이 새로운 Ler regulon인 Z0395 유전자를 포함하고 있음을 발견했다. Stx 모듈론은 Stx를 암호화하는 프로파지 유전자들 외에 추가로 *thi* 및 *cus* locus 유전자들을 포함했다. *thi* 및 *cus* locus 유전자들은 각각 티아민(thiamine) 및 구리이온에 의해 조절된다고 알려져 있기 때문에, Stx 프로파지 유전자들의 발현 역시 티아민 및 구리이온에 의해 조절됨을 분자생물학적 실험을 통해 확인했다. 따라서, ICA 모델을 통해 식별한 모듈론들은 장출혈성대장균의 성장 조건이나 유전적 배경에 관계없이 공동 조절되는 유전자들로 구성돼 있음을 확인했다. 또한, 모듈론들을 활용해 장출혈성대장균이 소를 감염 시키는 과정에서 필요한 독성 및 생존

유전자의 발현 조절을 성공적으로 설명할 수 있었다. 결과적으로, ICA 모델로 식별한 모듈론이 장출혈성대장균의 병원성에 중요한 유전자의 전사조절 네트워크를 포괄적으로 이해하고 확장시키는 데 기여할 수 있음을 보였다. 종합적으로, 본 연구는 장출혈성대장균의 대규모 전장유전체 및 전사체 데이터를 분석하는 기계학습 모델들을 개발했으며, 이를 통해 잠재적 병원성과 전사조절 네트워크를 조사하는 새로운 방법을 제안했다. 이러한 기계학습을 활용한 분석 방법은 신종 장출혈성대장균에의한 감염을 예방하고 대처하는 신기술이 될 수 있을 것이다.

핵심어: 장출혈성대장균, 기계학습, 유전체, 잠재적 병원성, 전사체, 전사조절 네트워크

학번: 2017-21443