저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 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: 

©Collection
Applications of linear and nonlinear models for association and classification using genomic polymorphisms

유전체 다형성을 이용한 연관과 분류를 위한 선형 및 비선형 모델의 적용

2019년 2월

서울대학교 대학원
협동과정 농생명유전체학전공
장 동 성
Applications of linear and nonlinear models for association and classification using genomic polymorphisms

유전체 다형성을 이용한 연관과 분류를 위한 선형 및 비선형 모델의 적용

지도교수 김희발
이 논문을 농학석사 학위논문으로 제출함
2018년 12월

서울대학교 대학원
협동과정 농생명유전체학전공
장동성

장동성의 농학석사 학위논문으로 인준함
2018년 12월

위원장 윤철희
부위원장 김희발
위원 조서애

서울대학교
SEUL NATIONAL UNIVERSITY
Abstract

Applications of linear and nonlinear models for association and classification using genomic polymorphisms

Dongsung Jang
Interdisciplinary Program in Agricultural Genomics
The Graduate School
Seoul National University

I performed associations and classifications to identify genomic variants from genomic polymorphisms. In the first study, Genome-Wide Association Studies (GWAS) were conducted to detect the association between SNP markers and phenotypes. For this process, I created linear models using the PLINK program. In the second study, I conducted decision tree for classifications using genomic regions to detect genomic variants that can classify species. Decision trees are nonlinear models and particularly applicable for multiclass dependent variables (Kim 2003; Ozaki 2014). In this study, the dependent variables were pathogenic *E. coli* strains, which were
classified into five types according to the symptoms.

In chapter 2, I performed multivariate GWAS analysis using SNP markers and phenotypes of Berkshire and Duroc. In general, Berkshire is more tender compared to Duroc. I conducted this study in order to detect what genomic variation causes these differences in pork quality. The target phenotype in this study was tenderness of the pork. Because this tenderness was a complex trait affected by various phenotypes, I performed multivariate GWAS rather than univariate GWAS. Based on the result of multivariate GWAS, I could identify SNP marker (ASGA0033314) that were related to tenderness and suggested candidate genes (MEP1A, ARPC1A, ENAH, APC2) related to this marker.

In chapter 3, I performed classification analysis to classify pathogenic E. coli. Among the pathogenic E. coli, I used five types of E. coli strains related to diarrheal disease known to cause food poisoning in this study. In general, pathogenic E. coli is classified into the presence or absence of virulence factors. This study was conducted to propose different classification method different from the existing one using sliding window technique. I conducted decision tree analysis to detect genomic regions that could classify pathogenic E. coli. The genomic regions detected by decision tree is presented as regions
in which genomic variations exist to classify *E. coli*. Among these regions, there are sequences that have not yet been studied. If the functions of these regions are studied, they are expected to be used as markers or primers. This study is expected to contribute to the treatment of pathogenic *E. coli* related symptoms by the classification of pathogenic *E. coli* by the sliding window method.

**Key words:** genomic polymorphisms, multivariate GWAS, classification, genomic region, sliding window, decision tree

**Student number:** 2016-24132
Table of Contents

Abstract ....................................................................................................................... i
Table of Contents ...................................................................................................... iv
List of Figures ............................................................................................................ vi
List of Tables ............................................................................................................. viii
Chapter 1. Literature Review .................................................................................. 1
  1.1 Overview ............................................................................................................ 2
    1.1.1 Linear and Nonlinear .................................................................................. 2
  1.2 Genome-Wide Association Study (GWAS) .................................................... 4
    1.2.1 Multivariate analysis .................................................................................. 7
    1.2.2 Multiple comparison problem .................................................................... 9
  1.3 Classification .................................................................................................... 10
    1.3.1 Escherichia coli ........................................................................................ 11
    1.3.2 Sliding window .......................................................................................... 12
    1.3.3 Cross-Validation ....................................................................................... 13
Chapter 2. Multivariate genome-wide association studies on tenderness of Berkshire and Duroc pig breeds ........................................... 16
  2.1 Abstract ......................................................................................................... 17
  2.2 Introduction .................................................................................................... 18
Chapter 3. Identification and classification criteria for genomic regions involved in various types of pathogenic *Escherichia coli* using sliding window

3.1 Abstract

3.2 Introduction

3.3 Materials and Methods

3.4 Results

3.5 Discussion

References
List of figures

Figure 1-1. The following is a comparison of univariate GWAS and multivariate GWAS analysis.................................................................8

Figure 1-2. The figure below is an example of “10-fold Cross Validation” for modeling and evaluating by dividing 10 clusters from given datasets........15

Figure 2-1. Flow chart for multivariate genome-wide association studies.21

Figure 2-2. A network plot showing the correlation of phenotypic traits presented in the Table 2-1. The degree of correlation depends on the color of the line. (a) shows the correlation between Berkshire traits and (b) shows the correlation between traits of Duroc.................................................................24

Figure 2-3. The Manhattan plot (left) and QQ-plot (right) of SNPs resulting from MV-GWAS with CieL and Heat-loss. The dotted line is the threshold to solve the multiple comparison problem—the blue line is the Bonferroni correction, and the red line is the FDR-BH. The ASGA0033314 means significant SNP from interaction effect of MV-GWAS .................................32

Figure 3-1. Flow chart for classification of E. coli used in this study........44

Figure 3-2. Concept of Sliding Window Analysis. The alignment score shown in this figure are examples ..........................................................54

Figure 3-3. The number of 5 pathogenic E. coli samples was used in this study ..........................................................63
Figure 3-4. Concept of data transformation used in our analysis ............65

Figure 3-5. These matrices count how many samples that truly belong to each type (columns) were predicted to belong to that type (rows) when we set the type criteria differently .................................................................71

Figure 3-6. The graph below is a harmonic average that calculated Matthews Correlation Coefficient (MCC) values by type (EHEC, EPEC, and ETC). If the MCC is not calculated, no value is displayed in the graph..................72

Figure 3-7. Decision tree model and confusion matrix when window size is 5,000 bp, overlap size is 500 bp and the number of features is 400.........74
List of Tables

Table 2-1. Summary statistics for meat quality traits in a Berkshire and Duroc population ................................................................................................................. 23

Table 2-2. KEGG pathways from gene lists in which significant SNPs are located by MV-GWAS ........................................................................................................... 33

Table 2-3. Candidate genes that contribute to tenderness of pork in Berkshire pigs .......................................................................................................................... 36

Table 3-1. Description of the different datasets used ........................................ 47

Table 3-2. Major virulence factors from VFDB in E. coli related to Pathogenic E. coli .................................................................................................................. 51

Table 3-3. Different modalities of window size and corresponding overlap size used ...................................................................................................................... 55

Table 3-4. Number of features according to window and overlap size ........... 66

Table 3-5. Example of data transformation by sliding window analysis ......... 67

Table 3-6. Example of the importance of features ........................................... 69
Chapter 1. Literature Review
1.1 Overview

I conducted research to detect variations from the genome in the population. It is Single Nucleotide polymorphism (SNP) that determines the identify of an individual, which is a nucleotide sequence that occurs every 1,000 base pairs (bp) in the DNA sequence. I used SNP and genomic regions, which are variations representing individual characteristics, to analyze the association with phenotypes from SNP and to classify individuals using genomic region. Therefore, I could understand the tendency by discovering variations that represent individual characteristics.

1.1.1 Linear and Nonlinear

A term "linear" in linear model refers to the linear in parameters, meaning that there is a linear relationship with respect to the parameters. As equation (1) shows, the linear model has the dependent variable $y$, the independent variable $x_i$, and the parameters $\beta_i$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad \cdots (1)$$

If $x_i$ is a function of only independent variables, the above equation is the
linear model. In many cases, the functional relationship between $y$ and $x_i$ is not linear. The nonlinear model is the nonlinear relationship of $y$ and $x_i$ to the parameters. For example,

$$y = \exp(\beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon) \quad \cdots \quad (2)$$

is nonlinear relationship to the parameters.

However, the two equations above are fundamentally different. If conduct the logarithm transformation both equation (1) and equation (2), equation (1) can be transformed into the linear model for the parameters $\beta_i$ as

$$\ln y = \beta_1 + \beta_2 x + \beta_3 x^2 + \epsilon.$$

What can be transformed into the linear model through this transformation is called intrinsically linear.

On the other hand, equation (2) cannot be linearized with respect to the parameter $\beta_i$ regardless of the transformation, and such a model is called intrinsically nonlinear. The linear models are simple to interpret for parameter coefficients, but nonlinear models are very difficult to interpret if the model is complex. Therefore, nonlinear models are not used well in statistical modeling, which emphasizes the interpretation of models.

However, if the purpose of the model is predictive rather than interpretive, the nonlinear model is very flexible, so it is possible to model data with
complex patterns. So if you have enough datasets to reduce the error of variance sufficiently and the prediction itself is your goal, then the nonlinear model is a very useful tool. In the field of machine learning, this nonlinear model is actually used. To summarize, the linear model is the model in which parameters are represented in linear forms. And these linear models are relatively easy to estimate parameters or to interpret compared to nonlinear models. If you can create a linear model by appropriately transforming data or adding meaningful features, this is more advantageous than creating a complex nonlinear model with a small number of features. However, linear models are less flexible because they are limited in the type of combination of parameters.

Therefore, accurate modeling of data with complex patterns may not be possible. In recent, nonlinear models have been widely used in areas where sophisticated prediction is more important than model interpretation.

1.2 Genome-Wide Association Study

Recent genetic studies have focused on understanding the mechanism of its expression by identifying phenotypes associated with each gene location.
Genome-Wide Association Study (GWAS) is an exploratory method of finding phenotypes associated with such genes. GWAS is an analysis that finds genetic locations that are related to a target phenotype for all genomic locations. In fact, it is really difficult to find out experimentally how many genes are associated with a number of genes. Because GWAS analyzes association for all gene locations, it is a very useful search tool for finding candidate genes that are primarily involved in a phenotype or disease of interest. However, even if the gene is found in GWAS, it is not always the causative gene. In other words, GWAS is not a search for a causal relationship but a process of finding candidates for genes that appear incidentally. Thus, in general, the study proceeds with a search for candidate genes through GWAS, followed by the results of biological validation in more animal and cell experiments, and finally the identification of gene-phenotype relationships. GWAS must be a powerful tool, but it is important to clearly recognize the limitations because the principles are due to statistical association analysis.

In addition, there is an important concept to understand the GWAS analysis method, ‘Linkage Disequilibrium (LD)’. We inherit a pair of genes from the
parent, which divides the reproductive cells and causes a permanent genotypic rearrangement in the same cell. However, since genetic recombination occurs in a block, the genotypes that are located close to each other are not mixed together but moved together in a mosaic pattern. One such block is commonly referred to as an ‘LD block’. For the positions included in the same LD block, the correlation is analyzed and the p-value showing the same correlation are shown. The presence of the LD block suggests the following four;

1) GWAS does not have to do for every gene location. Only one representative marker in the same LD block can be used. That is, the number of analysis positions is simplified.

2) Even if the candidate position is found by GWAS, the precise cause gene position may be another position in the LD block. After GWAS, all genotypes near that location should be reviewed again in detail.

3) In the Manhattan plot, which is commonly used in GWAS, the signal is all around the top, like a tower.

4) It is possible to estimate the genotype of an unchecked region in the same LD block, often called imputation.
1.2.1 Multivariate analysis

The quality of pork is a complex trait that is influenced by various phenotypes. In particular, the degree of tenderness is one of the most important quality factors in pork. GWAS is an analysis based on the premise that genetic variants are responsible for the variety of phenotypes. Among them, GWAS finds genetic variants in single nucleotide polymorphism. The target trait of this GWAS is the tenderness of meat, which is a complex trait involving various traits. Therefore, we conducted multivariate GWAS for various traits (multivariate) rather than one trait (univariate).

Multiple traits, that is, multivariate analysis, can be very advantageous compared to monolayer analysis for a variety of reasons. First, multivariate analysis increases the validity of genetic correlations between different genetic traits. Additional information provided by cross-trait covariance is ignored in the univariate analysis. Second, most multivariate procedures can perform a single test on the association with a set of traits. This reduces the number of tests performed and reduces the multiple test burden compared to analyzing all the characteristics individually.
**Figure 1-1.** The following is a comparison of univariate GWAS and multivariate GWAS analysis. (a) and (b) are the manhattan plots resulting from the univariate GWAS and (c) is the manhattan plots resulting from the multivariate GWAS used in the traits used in (a) and (b).

(a)

(b)

(c)
1.2.2 Multiple comparison problem

GWAS usually utilizes 500,000 ~ 2,500,000 SNPs for analysis, so multiple comparison problems must occur. If the significance level of the test using one SNP is called as $\alpha$, the result of one test is wrong with a probability of $\alpha \times 100\%$. If you test with 500,000 SNPs, the probability that you will make a false conclusion even once, that is, false positive rate is $1 - (1 - \alpha)^{50,000}$, which contains about 100% errors. Various statistical techniques have been devised to overcome these problems. In this study, I used the “Bonferroni correction’ and “FDR-BH” method.

Bonferroni correction based on Bonferroni’s inequality is a method of dividing the significance level of the whole analysis by the number of all the tests and making it a significance level in one test. That is, if the number of SNPs used in the analysis is $k$, the significance level $\alpha$ used in the one-time test is $\frac{\alpha}{k}$, and the significance probability smaller than the newly obtained $\alpha$ is a method of recognizing significance only for the observed SNPs. This method has the advantage of simple calculation. However, since all SNPs are independent, that is, assuming Linkage Disequilibrium (LD), there is a disadvantage that the level of significance used in one test becomes too small.
and the power is reduced. In other words, there is a greater probability of making errors that are not relevant even if there is a relationship between the phenotype and the SNP. There are several alternatives to improve the problem of the Bonferroni correction. Among them, there is an approach using FDR that suppresses the ratio of false judgments among SNP judged to be related to the phenotypes to be less than a certain ratio. In particular, FDR-BH method judges that SNPs having a p-value smaller than the j-th small significance probability $p_j$ are related to each other in order of significance from each test. In this case, when the significance level of the entire analysis is $\alpha$, it is evaluated from the large p-value using the relation of $p_j \leq \alpha \times j/n$, and when the first inequality is established, the SNPs having a value are judged to be significant.

1.3 Classification

Classification is a kind of supervised learning, in which the category of existing data is identified and the category of newly observed data is determined by itself. By learning existing categories that contain a set of data, and based on this, computers learn to divide boundaries by separating
categories of data. Therefore, new data entered the model are classified according to the close category or learned algorithm depending on where the data is located. The classification algorithm used in this study is Decision Tree (DT). One of the simplest classifiers, it has a very simple structure that draws models graphically using tools. This is the model that will make the final decision while choosing the significant node from root. The advantage of this tree is that anyone can easily understand it, so it is easy for anyone to use.

1.3.1 *Escherichia coli*

*Escherichia coli* (*E. coli*) has various properties depending on the serotype, and *E. coli* is classified into 180 kinds of O antigen, 90 kinds of K antigen and 50 kinds of H antigen by the immunological specificity of the antigen component constituting the cell. The morphological characteristics of *E. coli* are gram negative, and it has motility due to the main maternal flagella, but there are also no flagella and non-motility. Also, biochemical characteristics are aerobic or facultative anaerobic bacteria producing acid and gas by decomposing Lactose and Fructose. The growth temperature is 7 ~ 48 °C,
the optimal temperature is 35 ~ 37 °C, and the growth optimal pH range is 4.5 ~ 9.0.

*E. coli*, which is associated with diarrheal disease, is a major enterohemorrhage *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteropathogenic *E. coli* (EPEC) enteroinvasive *E. coli* (EIEC), and enteroaggregative *E. coli* (EAEC). The aim of this study was to identify these five pathogenic *E. coli* as genomic regions that can be newly classified, rather than classifying them as existing antigen or virulence factors. Using the entire *E.coli* genome, it is the significance of this study to find a variation regions (as a genomic regions) that can classify these pathogenic *E. coli* that could not be found previously.

### 1.3.2 Sliding window

Sliding window analysis is a method of determining a certain window size when comparing sequences and comparing the sequences to check whether or not the sequences of the corresponding region exceed the similarity cutoff. This method helps to identify only significant regions of similarity between
sequences. Of course, window sizes and thresholds can be set differently according to the degree of similarity between the sequences to be compared, and it is best to determine two values by trial and error when analyzing new data. The sliding window method was applied to identify genomic regions to classify *E. coli*. I divide the genome into a specific size, name it a window (or genomic region), and move it at regular intervals to find the sequence differences with other samples. If there is a difference in the type of *E. coli*, the genomic region is presented as a genomic region.

### 1.3.3 Cross validation

Cross validation is a method of evaluating a model in statistics. The basic method for evaluating the model is to separate training datasets and test datasets from the given data. Then create a model with the training datasets and evaluate the model's performance with the test datasets. The cross validation is used to overcome the problems of this basic method. The problem is that if the size of the dataset is small, the performance of the test set is not reliable. If performance is very different depending on how you set up the test datasets, the effect of chance will bias the model evaluation index.
To overcome this problem, cross validation allows all data to be used at least once in test datasets.

In this study, “K-fold Cross Validation” is performed. This method randomly divides the datasets into “K” folds. Then, one of each group becomes test datasets for validation, and the rest becomes training datasets and performs learning for the model. As a result, the test datasets are evaluated “K” times based on the learned model from training datasets.
**Figure 1-2.** The figure below is an example of “10-fold Cross Validation” for modeling and evaluating by dividing 10 clusters from given datasets.
Chapter 2.
Multivariate genome-wide association studies on tenderness of Berkshire and Duroc pig breeds
2.1 Abstract

Genome-wide association studies (GWAS) have been steadily used for identification of genomic links to disease and various economical traits. Of those traits, a tenderness of pork is one of the most important factors in quality evaluation of consumers. In this study, we use two pig breed populations; Berkshire is known for its excellent meat quality and Duroc which is known for its high intramuscular fat content in meat. Multivariate genome-wide association studies (MV-GWAS) was executed to compare SNPs of two pigs to find out what genetic variants occur the tenderness of pork. Through MV-GWAS, we have identified candidate genes and the association of biological pathways involved in the tenderness of pork. From these direct and indirect associations, we displayed the usefulness of simple statistical models and their potential contribution to improving the meat quality of pork. We identified a candidate gene related to the tenderness in only Berkshire. Furthermore, several of the biological pathways involved in tenderness in both Berkshire and Duroc were found. The candidate genes identified in this study will be helpful to use them in breeding programs for improving pork quality.
### 2.2 Introduction

A pork industry has been steadily growing with the agricultural advances, such as statistical genetics and molecular breeding. A tenderness of pork is one of the most important factors in the quality evaluation of consumers. This pork quality is a complex trait influenced by a combination of genetic and non-genetic factors. Several studies have revealed the genetic mechanisms behind it (Rosenvold and Andersen 2003). So, the tenderness varies due to not only the genetic factors but also the non-genetic factors such as breeding conditions. This is why a variation in the chemical composition or structure of the muscle fibers and connective tissue causes various differences. These studies illustrate the direct relationship between DNA markers and pork quality - it is affected by several genes that are mapped to the quantitative trait loci (QTL) regions. To assess the genetic association to such complex traits, researchers focused on mapping and characterization of trait loci, and farm-animal genome analysis (Andersson 2001, Andersson and Georges 2004). The genetic variants between and within breeds are thought to hold the key to understanding the variation in meat quality traits such as tenderness and intramuscular fat content (IMF). By neglecting the environmental factors
that cannot be measured, the importance of DNA marker (or single nucleotide polymorphism; SNP) investigation has been proved more than a decade ago (De Vries, Faucitano et al. 2000).

The Genome-wide association study (GWAS) has been taken in as the golden standard or primary method for investigating the association between genetic variants and phenotypic traits (both quantitative and qualitative) in Humans and farm animals. With this regard, several studies dealt with meat quality traits (Luo, Cheng et al. 2012), based on both single trait and/or multiple trait analysis through phenotype integration (Lee, Shin et al. 2014). Such phenotype integration may cloud the water in terms of interpretation; the findings from such analyses are perhaps less obvious when the dependent variable is created from a combination of factors. We focused on the interpretation of the results where the association between a genetic variant and the trait is evident. By using multivariate models for pre-screening and interpreting their results with posthoc analyses through univariate models can provide powerful candidates of interest, where a good proportion of those markers have been validated through references of genome and transcriptome studies. The employed methods, including reference-based validation, are
widely used and have already displayed their successfulness (Galesloot, Van Steen et al. 2014, Iqbal, Kim et al. 2015); it is, therefore, practical to accept the novel findings from our study to be noteworthy.
**Figure 2-1.** Flow chart for multivariate genome-wide association studies.
2.3 Materials and Methods

Data Preparation

We compared the species-specific and global variants—significant variants in both species—using 60K chip data of Berkshire and Duroc breeds. Both of the breeds are known for their superior meat quality, and thus our results may provide keys to improving eating qualities of pork according to National Pork Producers’ Council (NPPC) Terminal Sire-Line Genetic Evaluation Program (Goodwin and Burroughs 1995). There were a total of 695 Berkshire (Male: 361, Female: 201, castration: 133) and 157 Duroc (Only female) pig breeds. We used phenotypic traits of back fat thickness (BF), meat color (specifies as L*, A*, B* color scale) (Davis and Ohno 2010), carcass weight (CWT), drip loss (fluid lost from fresh, non-cooked meat via passive exudation), intramuscular fat content (IMF), heat loss (amount of meat decreases when heated at specific temperature), moisture, meat pH 24h post-mortem (OnedaypH), protein content, shear force (SF), water holding capacity (WHC, the ability of meat to retain its inherent moisture even through external pressures). L* is a measure of lightness where 0 equals black and 100 equals white. A* indicates redness and greenness, and B* indicates yellowness and blueness. Table 2-1 presents a description of traits analyzed.
**Table 2-1.** Summary statistics for meat quality traits in a Berkshire and Duroc population.

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Berkshire (n=695)</th>
<th>Duroc (n=157)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>CV</td>
</tr>
<tr>
<td>Backfat</td>
<td>25.16 ± 0.39</td>
<td>21</td>
</tr>
<tr>
<td>CieA</td>
<td>6.18 ± 0.075</td>
<td>16</td>
</tr>
<tr>
<td>CieB</td>
<td>3.23 ± 0.09</td>
<td>37</td>
</tr>
<tr>
<td>CieL</td>
<td>48.53 ± 0.21</td>
<td>6</td>
</tr>
<tr>
<td>CWT</td>
<td>85.94 ± 0.405</td>
<td>6</td>
</tr>
<tr>
<td>Driploss</td>
<td>4.42 ± 0.14</td>
<td>41</td>
</tr>
<tr>
<td>Heatloss</td>
<td>26.08 ± 0.32</td>
<td>3</td>
</tr>
<tr>
<td>IMF</td>
<td>2.82 ± 0.085</td>
<td>17</td>
</tr>
<tr>
<td>Moisture</td>
<td>75.37 ± 0.08</td>
<td>1</td>
</tr>
<tr>
<td>OnedaypH</td>
<td>5.78 ± 0.015</td>
<td>43</td>
</tr>
<tr>
<td>Protein</td>
<td>23.96 ± 0.06</td>
<td>3</td>
</tr>
<tr>
<td>SF</td>
<td>2.88 ± 0.05</td>
<td>24</td>
</tr>
<tr>
<td>WHC</td>
<td>59.12 ± 0.23</td>
<td>5</td>
</tr>
</tbody>
</table>

CV: Coefficient of Variation, CV = (Standard Deviation/average) * 100 (%); CIE, Commission Internationale de l’Eclairage.
Figure 2-2. A network plot showing the correlation of phenotypic traits presented in the Table 2-1. The degree of correlation depends on the color of the line. (a) shows the correlation between *Berkshire* traits and (b) shows the correlation between traits of *Duroc*.

(a) 

(b)
Genome-Wide Association Studies

Multivariate GWAS

The study was carried out to understand the relationship between pork quality traits and genotype analysis using a Porcine 60K beadchip for Berkshires and Duroc breeds. We analyzed information from 62,163 SNPs in Berkshire pigs and 44,345 SNPs in Duroc pigs from porcine 60K beadchip analysis kit (Illumina, USA). SNPs data was used for multivariate association and linear regression analysis using multivariate PLINK and PLINK version 1.9, respectively.

In our GWAS analysis, we used only 18 autosomes. The 62,163 (Berkshire) and 44,345 (Duroc) of Porcine 60K beadchip SNP data was filtered under specific filtering conditions of: Minor allele frequency (MAF) <0.05, genotyping error > 0.01, and Hardy Weinberg equilibrium (P-value ≤ 0.001).

We used multivariate GWAS instead of univariate GWAS since multivariate GWAS considers interaction effect from two phenotypic traits in the analysis as compared to univariate GWAS that considers only one phenotypic trait. In other words, multivariate GWAS uses much more information to identify the differences between phenotypic traits as compared to univariate GWAS.
PLINK program was used for the analysis using the following multivariate linear regression equation.

\[ Y_h = X\beta_h + e_h \quad (1) \]

\[ h = 1,2 \quad e_h \sim N(0,I\sigma_h) \]

**Univariate GWAS**

When multivariate GWAS is performed, it cannot be determined whether the SNPs are derived from phenotype which was highly correlated or from interaction effect. In order to clarify the interaction effect through multivariate GWAS, an additional univariate GWAS is performed and the results are compared to the multivariate GWAS results. To find the only interaction effect from multivariate GWAS results, the SNPs from the univariate GWAS result should be eliminated if the relevant SNP from the multivariate GWAS results is derived from the relevant SNP from the univariate GWAS result.
Covariate adjustment and Residual analysis

We tested the appropriateness of a linear regression model by defining residuals that the linear model is not always appropriate for the data. Since the model is strongly influenced by an outlier, it is necessary to confirm whether or not the outlier intervenes or not. In Multivariate PLINK\(^1\), covariate option and multivariate analysis do not work together. Before MV-GWAS, we have to adjust our phenotypes respectively for any covariates using common statistical packages (eg. R). We used sex and age of the animal at slaughter as a covariate in this study. An Estimation of \( \beta_h \) in model (1) can be obtained by its BLUE (Best Linear Unbiased Estimation);

\[ b_h = (X'X)^{-1}X'Y_h \quad (2) \]

The residual vector is obtained by (2) using ordinary least squares (OLS) estimator; residual vector:

\[ e_h = Y_h - X_h b_h \quad (3) \]

---

**Multiple comparison problems**

A Multiple comparison problems occurs when one considers a set of statistical inferences simultaneously or infers a subset of parameters selected based on the observed values. When several dependent or independent statistical tests are performed simultaneously, the correction is used. To solve this problem, we used Bonferroni correction for the threshold. The Bonferroni correction is one of the several methods used to counteract this problem. In order to eliminate false significance, the alpha value needs to be lowered considering the number of comparisons being performed (Weisstein 2004).

**Gene annotation and KEGG pathway**

The *Ensembl Sscrofa build 10.2*[^2] was used to determine which genes are located in the significant SNPs in the MV-GWAS. SNPs should cover the gene region as well as its 30 Kb 5’ upstream and 30 Kb 3’ downstream regions (Wang, Liu et al. 2005). We identify whether the function of these genes is related to the pork quality.

Also, the *DAVID 6.8 functional annotation bioinformatics microarray*[^2]

analysis (https://david.ncifcrf.gov/) is carried out using gene list by gene annotation of significant SNPs to find out the KEGG pathway related to pork quality. Genes associated with pork quality are presented as candidate genes.

2.4 Results

**Genome-Wide Association Studies**

Regardless of correlation between phenotypic traits, we used 78 dependent variables on multivariate GWAS selecting 2 of 13 phenotypes in order to consider interaction effect by multivariate analysis. As a result of the analysis, significant SNPs by bonferroni correction were found. However, the main objective of this study is to find SNPs due to the interaction effects of phenotypes by multivariate analysis. Thus, we excluded SNPs by univariate GWAS from MV-GWAS. We did MV-GWAS only with SNPs by interaction effects.

Considering the genomic inflation factor of QQ-plot in both Berkshire and Duroc, we found significant SNP in the interaction effect between the two traits with CieL and Heat-loss in MW-GWAS of only Berkshire (**Figure 2-2**). This significant SNP, ASGA003314, was annotated with *Ensembl Sscrofa*
build 10.2. Then we found *MEP1A* (Meprin A Subunit Alpha) at 30 Kb 3’ downstream from SNP location. This *MEP1A* is associated with several biological pathways. Among these pathways, we focus on ‘collagen chain trimerization’. Collagen is an abundant connected tissue protein and affects variation in meat tenderness. Especially, this collagen plays an important role in cooked meat (Weston, Rogers et al. 2002).

**KEGG pathway**

Without considering the genomic inflation factor of the QQ-plot, we can annotate genes from more significant SNPs. Through this gene annotation process, multiple gene lists were obtained for each of the multivariate phenotypes. The functional annotation tool was used in DAVID 6.8 to find the KEGG pathways associated with these gene lists.

Several KEGG pathways were resulted from the analysis. Of these pathways, we were interested in the pathway that has been found to be related to the IMF from previous studies (Hamill, McBryan et al. 2012). Previous studies have shown that IMF in pork affects tenderness (Van Laack, Stevens et al. 2012).

---

3 http://pathcards.genecards.org/card/collagen_chain_trimerization
The KEGG pathways associated with the IMF of pork identified in the study are presented in **Table 2-2**. These pathways are related to IMF (Hamill, McBryan et al. 2012).
**Figure 2-3.** The Manhattan plot (left) and QQ-plot (right) of SNPs resulting from MV-GWAS with CieL and Heat-loss. The dotted line is the threshold to solve the multiple comparison problem - the blue line is the Bonferroni correction, and the red line is the FDR-BH. The ASGA0033314 means significant SNP from interaction effect of MV-GWAS.
Table 2-2 KEGG pathways from gene lists in which significant SNPs are located by MV-GWAS.

<table>
<thead>
<tr>
<th>Breed</th>
<th>Term</th>
<th>Multivariate phenotypes</th>
<th>P-value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkshire</td>
<td>ssc04810: Regulation of actin cytoskeleton</td>
<td>CieB and Driploss</td>
<td>0.099554511</td>
<td><strong>ARPC1A, ENAH, APC2, ITGAV, FGF10</strong></td>
</tr>
<tr>
<td>Berkshire</td>
<td>ssc04530: Tight junction</td>
<td>CieB and OnedaypH</td>
<td>0.022446733</td>
<td><strong>JAM3, AKT3, PRKCB, CTNNB1</strong></td>
</tr>
<tr>
<td>Ssc00190: Oxidative Phosphorylation</td>
<td></td>
<td></td>
<td>0.042659674</td>
<td><strong>NDUFS7, SDHB, ATP6V1G3, NDUFS2</strong></td>
</tr>
<tr>
<td>Duroc</td>
<td>Galactose metabolism</td>
<td>Backfat and CWT</td>
<td>0.023316812</td>
<td><strong>B4GALT1, GANC, GAA</strong></td>
</tr>
</tbody>
</table>

Up-regulated genes are highlighted in bold, down-regulated genes are in italics.
Candidate gene

From multivariate GWAS results, we obtained several of genes from significant SNPs. Analyzing the function of these genes, some genes could directly or indirectly affect tenderness of pork in Berkshire. *MEPIA* (Meprin A Subunit Alpha) is related to collagen. *ARPC1A* (Actin-Related Protein 2/3 Complex Subunit 1A), *ENAH* (ENAbled Homolog), and *APC2* (Adenomatous Polyposis Coli Protein 2) are related to actin. These genes are candidate genes that contribute to the tenderness of pork in Berkshire pigs.

Pork quality is a complex trait affected by both genetic and non-genetic factors. Pork quality traits including tenderness, IMF, color, and drip-loss are polygenetic that more than one gene affects a trait and traits themselves are affected each other. Tenderness is among the most important quality parameter of pork that is affected by several factors. The contribution of IMF of pork to tenderness and color have been investigated previously. This study investigated candidate genes that are associated with tenderness of pork. The genes associated with tenderness found in this study may be helpful in improving the meat quality of pork. From these direct and indirect associations, we displayed the usefulness of simple statistical models and
their potential contribution to improving the meat quality of pork.
### Table 2-3. Candidate genes that contribute to tenderness of pork in Berkshire pigs.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MEP1A</strong></td>
<td>It is related to the pathway of Collagen chain trimerization.</td>
</tr>
<tr>
<td><strong>ARPC1A</strong></td>
<td>It is involved in regulation of actin polymerization</td>
</tr>
<tr>
<td><strong>ENAH</strong></td>
<td>It is actin-associated proteins involved in a range of processes dependent on cytoskeleton remodeling</td>
</tr>
<tr>
<td><strong>APC2</strong></td>
<td>It stabilizes microtubules and may regulates actin fiber dynamics</td>
</tr>
</tbody>
</table>
2.5 Discussion

Tenderness of pork is one of the most important quality factors in meat and is a complex trait that is influenced by various traits. We conducted multivariate GWAS rather than univariate GWAS to present candidate genes from SNPs that affect these complex traits. *MEP1A* was associated with collagen, *ARPC1A, ENAH*, and *APC2* were associated with actin. These candidate genes could help improve the meat quality of pork.
This chapter will be published in elsewhere as a partial fulfillment of Dongsung Jang’s Master program.

Chapter 3.
Identification and classification criteria for genomic regions involved in various type of pathogenic *Escherichia coli* using sliding window
3.1 Abstract

Pathogenic *E. coli*, which can cause diseases such as diarrhea, is classified as five types (EAEC, EIEC, EHEC, ETEC, EPEC), each type being associated with specific virulence factors. As the treatment depends on the type of *E. coli* infecting the patient it is essential to classify the pathogenic *E. coli* rapidly and precisely. This study aims to find novel effective and quick ways to classify these pathogenic *E. coli*. Study was conducted in three steps; 1) *E. coli* data collection, 2) data transformation using alignment and sliding window analysis, 3) classification analysis using decision tree. Decision tree analysis is the supervised learning analysis method that presents the advantage of being intuitively understood and easily explained.

Also, since various decision tree algorithm packages have been developed in R, it is possible to compare the prediction accuracy of these various methods and to select the package with the best prediction performance. In this analysis, we used the “party” package. This is because the party package that uses the P-value of the nodes in the pruning method was more convenient to use than the other packages that use pruning using the new index. As a result, the classification model issued from the “party” package gave the
highest prediction accuracy and Matthews correlation coefficient and genomic regions classifying the pathogenic E. coli were identified. These genomic regions are now available as a novel way to classify pathogenic E. coli.

3.2 Introduction

Big data has been growing exponentially in biology and particularly in the field of genomics since Next Generation Sequencing (NGS) technology was introduced in the mid-2000s. Classifying groups of organisms based on genomic data using statistical programs such as R has thus become common practice. Classification analysis is a technique consisting in assigning observed data to groups predetermined by using a model calibrated on data belonging to already identified groups. A major form of classification analysis is decision tree Smith and Tansley (2004) and will be used in this study to classify microorganisms.

Microorganisms or microbes constitute a substantial component of the ecosystem. While most microbes are beneficial either at the biotope such as decomposition of organic matter or at the organism scale such as probiotic,
some of them, known as pathogenic microorganisms, can cause various
diseases such as food poisoning and sepsis so that humans have long been
making strategies to contain their outburst. Pathogenic microorganisms
usually enter human body through the ingested food and can subsequently
trigger variable health concerns. Control over pathogenic microorganisms has
thus emerged as a major interest in food industry in order to avoid any threat
to human health. As identifying unclassified microorganisms is prerequisite
to treatment of these pathogenic microorganisms, and many of them are
currently classified using genotyping or using the presence or absence of
specific pathogens (Versalovic and Lupski 2002).

One of the major microorganism at the origin of food poisoning in human is
the pathogenic *Escherichia coli (E. coli)*. The different strains of pathogenic
*E. coli* have been to date classified by three means: by its antigens (O-, K-, H-, and F-), by its serotype according to the immune response of a specific
antibody to each antigen (Tamura, Sakazaki et al. 1996) and finally according
to the five types of virulence factors which have persisted to become specific
pathotypes (Kaper, Nataro et al. 2004): Enterotoxigenic *E. coli* (ETEC),
Enteroinvasive *E. coli* (EIEC), Enteropathogenic *E. coli* (EPEC),
Enterohemorrhagic *E. coli* (EHEC), Enteroaggregative *E. coli* (EAEC). The different types of pathogenic *E. coli* can cause various health concerns (Evans Jr and Evans 1983). In particular some strains can cause sepsis and pyelonephritis (Ohno, Ota et al. 2003). Hence, it is necessary to classify pathogenic *E. coli* rapidly and precisely in order to treat the symptoms efficiently.

Different types of *E. coli* strains may share common genes (Lu, Jin et al. 2016) and genes may also overlap each other so that classifying pathogenic *E. coli* strains by the presence or absence of specific genes might not lead to the most accurate results. To remedy these shortcomings, this study aims to propose a new method to classify *E. coli* using genetic variation in the *E. coli* genome without confirming the existence of specific genes. Classification analysis is an analytical method building a classification criteria used for predicting categorical response variables using categorical or continuous explanatory variables.

Decision Tree (DT) analysis is one of the most widely used methods of machine learning thanks to its characteristics: DT is computationally fast,
makes no statistical assumptions (Pal and Mather 2003) and is visually readable compared to other classification analyzes.

Here, first, a new classification method for the various types of pathogenic E. coli based on DT analysis will be proposed. Second an attempt to identify specific the genomic regions to each E. coli types using DT analysis will be made.
Figure 3-1. Flow chart for classification of *Escherichia coli* used in this study
3.3 Materials and Methods

A total of 82 samples of pathogenic *E. coli* issued from the Korean Ministry of Food and Drug Safety (Korea, Republic) were used for classification analysis. These datasets have been uploaded to the FORCDB website (http://forcdb.snu.ac.kr/). Additional public datasets of *E. coli* were downloaded from the GenomeTrakr Network to compensate the small effective of the initially available datasets and to obtain reliable analysis results. These data can be obtained from the National Center for Biotechnology Information (NCBI), each contributing labs of the GenomeTrakr Network having a unique accession number registered within the NCBI and the data in each lab having a Short Read Archive (SRA) number. Sequence data used in this study was downloaded directly from this SRA. After data collection, the total number of samples initially prepared for classification analysis was 2,722 (Table 3-1).

The *E. coli* strain Nissle 1917 (EcN) was used as a reference genome (Accession number: PRJNA248167). This strain is the active principle of a probiotic preparation (trade name Mutaflor®) used for the treatment of
patients with intestinal diseases such as ulcerative colitis and diarrhea (Reister, Hoffmeier et al. 2014).
Table 3-1 Description of the different datasets used.

<table>
<thead>
<tr>
<th>Contributing Labs</th>
<th>Accession</th>
<th># of samples</th>
<th>Contributing Labs</th>
<th>Accession</th>
<th># of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ministry of Food and Drug Safety (Korea, Republic)</td>
<td>-</td>
<td>82</td>
<td>Minnesota Department of Health</td>
<td>PRJNA311907</td>
<td>20</td>
</tr>
<tr>
<td>Alaska State Public Health Laboratory-Anchorage</td>
<td>PRJNA306263</td>
<td>5</td>
<td>New Mexico State University Food Safety Laboratory</td>
<td>PRJNA292570</td>
<td>45</td>
</tr>
<tr>
<td>Argentina</td>
<td>PRJNA282762</td>
<td>52</td>
<td>New York State Department of Health Wadsworth Center</td>
<td>PRJNA232925</td>
<td>17</td>
</tr>
<tr>
<td>Arizona State Public Health Laboratory</td>
<td>PRJNA230968</td>
<td>729</td>
<td>NSF International</td>
<td>Applied Research Center</td>
<td>PRJNA357822</td>
</tr>
<tr>
<td>California Department of Health - FDLB Micro</td>
<td>PRJNA277984</td>
<td>319</td>
<td>Ohio Department of Agriculture</td>
<td>Animal Disease Diagnostic Laboratory</td>
<td>PRJNA338676</td>
</tr>
<tr>
<td>FDA-CFSAN_IFSH</td>
<td>PRJNA358244</td>
<td>23</td>
<td>Penn State University</td>
<td>Department of Food</td>
<td>PRJNA357722</td>
</tr>
<tr>
<td>FDA-CFSAN MDP Escherichia coli survey from foods</td>
<td>PRJNA312475</td>
<td>163</td>
<td>SDSU Veterinary &amp; Biomedical Sciences/ DOH South Dakota</td>
<td>PRJNA299490</td>
<td>100</td>
</tr>
<tr>
<td>Florida Department of Health University of Florida</td>
<td>PRJNA298331</td>
<td>253</td>
<td>State Hygienic Laboratory at the University of Iowa</td>
<td>PRJNA338681</td>
<td>11</td>
</tr>
<tr>
<td>Joint Institute for Food Safety and Applied</td>
<td>PRJNA342435</td>
<td>157</td>
<td>Texas Department of State Health Services</td>
<td>PRJNA284275</td>
<td>105</td>
</tr>
<tr>
<td>Michigan Department of Agriculture and Rural Development</td>
<td>PRJNA368991</td>
<td>84</td>
<td>Thakur Molecular Epidemiology Laboratory NC State University</td>
<td>PRJNA293225</td>
<td>354</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2,722</td>
</tr>
</tbody>
</table>
Data Preparation

A total of 82 samples of pathogenic *E. coli* issued from the Korean Ministry of Food and Drug Safety (Korea, Republic) were used for classification analysis. These datasets have been uploaded to the FORCDB website (http://forcdb.snu.ac.kr/). Additional public datasets of *E. coli* were downloaded from the *GenomeTrakr Network* to compensate the small effective of the initially available datasets and to obtain reliable analysis results. These data can be obtained from the National Center for Biotechnology Information (NCBI), each contributing labs of the *GenomeTrakr Network* having a unique accession number registered within the NCBI and the data in each lab having a Short Read Archive (SRA) number. Sequence data used in this study was downloaded directly from this SRA.

After data collection, the total number of samples initially prepared for classification analysis was 2,722 (Table 3-1).

The *E. coli* strain Nissle 1917 (EcN) was used as a reference genome (Accession number: PRJNA248167). This strain is the active principle of a probiotic preparation (trade name Mutaflor®) used for the treatment of
patients with intestinal diseases such as ulcerative colitis and diarrhea (Reister, Hoffmeier et al. 2014).

**Data Processing**

**Grouping**

Unlike the MFDS datasets provided in 4 different groups (EAEC, EHEC, ETEC, EPEC) of *E. coli*, the *GenomeTrakr Network* provides datasets without grouping of *E. coli*. The genes representing each one of the five *E. coli* groups (EAEC, EIEC, EHEC, EPEC, ETEC) were thus used to assign each sample issued from the *GenomeTrakr Network* to these groups.

In this process, we identified unique virulence factors (Table 3-2) for each of the five pathogenic *Escherichia coli* in the VFDB\(^4\) and use Bowtie2 (Langmead, Trapnell et al. 2009).

Bowtie2 is a short read aligner that can take a reference genome and map single- or paired-end data to it (Trapnell and Salzberg 2009). It requires an indexing step in which one supplies the reference genome. Bowtie2 then creates an index that will be used in the subsequent steps to align the reads to

\(^4\) [http://www.mgc.ac.cn/VFs/](http://www.mgc.ac.cn/VFs/)
the reference genome (Schmeier 2018). Bowtie2 was used to group the
GenomeTrakr Network datasets into five *E. coli* based on these virulence
factors. The maximum fragment size (length of paired-end alignments +
insert size) was finally adjusted to 1000 base pairs (Schmeier 2018).
Table 3-2. Major virulence factors from VFDB in *E. coli* related to Pathogenic *E. coli*.

<table>
<thead>
<tr>
<th>Identified Virulence Factors</th>
<th>EAEC</th>
<th>ETEC</th>
<th>EIEC</th>
<th>EHEC</th>
<th>EPEC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adherence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAFs, Dispersin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adhesive fimbriae, EtpA</td>
<td></td>
<td></td>
<td>ECP, Efa-1/LifA, Intimin, Paa, ToxB</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Toxin</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAST1, Pet, Pic, ShET1</td>
<td></td>
<td>Heat-labile toxin (LT), Heat-stable toxin (ST)</td>
<td>ShET1, ShET2, Shiga toxin</td>
<td>Hemolysin, Stx</td>
<td>CDT, EAST1</td>
</tr>
<tr>
<td><strong>Iron uptake</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td>Aerobactin, Chu</td>
<td>Shu</td>
<td>-</td>
</tr>
<tr>
<td><strong>Protease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td>IcsP(SopA), Pic, SigA</td>
<td>EspP, SteE</td>
<td>EspC</td>
</tr>
<tr>
<td><strong>Regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td>Ler</td>
<td>Ler, Per</td>
</tr>
<tr>
<td><strong>Secretion system</strong></td>
<td></td>
<td></td>
<td>T2SS, TTSS</td>
<td></td>
<td>TTSS</td>
</tr>
<tr>
<td><strong>Type III translocated protein</strong></td>
<td></td>
<td></td>
<td></td>
<td>Cif, EspA, EspB, EspD, EspF, EspG, EspH, Map, NleA/EspI, NleC, NleD, Tir</td>
<td></td>
</tr>
<tr>
<td><strong>Actin-based motility</strong></td>
<td></td>
<td></td>
<td>IcsA(VirG)</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Alignment

Sequence alignment process is a prerequisite to comparing intergenic genomic sequences. The *E. coli* Nissle 1917 strain was used as a reference genome and paired-end alignment was performed using SAMtools (Li, Handsaker et al. 2009). As alignment results in SAM/BAM format files that are complex to read, a further step was performed to convert SAM files to the more easily readable BED format. The conversion utility bamtobed of bedtools (Quinlan 2014) was thus implemented as the additional data preprocessing step in the alignment process.

DATA Transformation

Sliding Window Analysis

Sliding window analysis consists in comparing similarities of alignment by sliding part of the sample sequence along a reference sequence (**Figure 3-2**). More specifically, when comparing the reference genome (**A**) with the sample genome (**B**) sequence, the size of the certain sequence region (which is referred to as “window”) is set in **B** and its sequence is compared to the sequence of **A** from the beginning to the end. The step size used when sliding
the window is referred to as “overlapping size”. For example, if the window size is set to 1,000 base pairs (bp) and slides by 10 bp, the similarities of the 1st to 1000th sequence of A are compared to the 1st to 1000th sequence of B.
**Figure 3-2.** Concept of Sliding Window Analysis. The alignment score shown in this figure are examples.
<table>
<thead>
<tr>
<th>Window size</th>
<th>Overlap size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 base pair</td>
<td>No overlap</td>
</tr>
<tr>
<td>5,000 base pair</td>
<td>500 base pair</td>
</tr>
<tr>
<td>5,000 base pair</td>
<td>No overlap</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>500 base pair</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>1,000 base pair</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>No overlap</td>
</tr>
</tbody>
</table>
Next, the 10th to 1010th sequence of A is compared with the 10th to 1010th sequence of B, and repeatedly until reaching the end of A and B sequences.

Window and overlap sizes were set as described in Table 3-3.

Sliding window analysis is performed to obtain an alignment score of the reference genome according to the position of the genome samples’ window. This alignment score means the similarities of the alignment. Here, the window’s alignment scores resulting from the sliding process were used as features (also called variables) in the classification model. For convenience of interpretation, these continuous alignment scores were coded into categorical features (1 or 0) in finding genomic variant regions in samples compared with the reference genome. The alignment score was coded as '1' if the similarities between sequences were greater than or equal to 0.9, and coded as '0' when the similarities were inferior to this 0.9 threshold.

**Feature Selection**

The more the features included in the analysis model, the greater the risk of overfitting the data. The purpose of feature selection is thus to reach the best predictive performances of the model by providing the deepest understanding
of the underlying processes with data as sparse as possible (Guyon and Elisseeff 2003). The features and responses considered in this study being both categorical, features were selected by implementing a Chi-squared test of independence (Kaushik 2017).

**Classification**

A classification analysis is an analytical method building a classification criteria for predicting categorical response variables using explanatory variables. Precisely categorized datasets of observations are needed for calibrating the classification criteria so that further new, undescribed observations can be accurately assigned in classification groups.

**Decision Tree**

A Decision Tree (DT) analysis, which is the most representative classification model, was conducted using the R software (Team 2013). Decision tree analysis is a nonparametric method expressing the criteria of classification as a tree structure. Particularly, trees with binary partition structure are widely used because they are easily understandable and interpretable, with reduced selection bias. The window for the location in the
sample was set as the explanatory variable and the five pathogenic *E. coli* were set as the response variable.

In several DT analysis packages, we will use “party package” in R. This package uses the recursive partitioning based on permutation tests method. The variables to be pruned are determined based on P-value of nodes so that the results are less likely to be biased and so that there is no need to further pruning. In particular, this method uses the P-value as a pruning criterion, so it is not affected even if the unit of the variables is different.

**k-Fold Cross Validation**

The basic method for evaluating a model is to separate the training datasets from the test datasets, then create the model with the training datasets and evaluate the performance with the test datasets (Kohavi 1995). Cross validation is to complement the problem of this basic problem. The problem is that if the size of the datasets is small, the performance of the test datasets is not reliable. If performance varies greatly depending on how you take the test datasets, the effect of chance will bias the model evaluation. To overcome
this, Cross validation ensures that all datasets are used to the test datasets at
least once.

We perform 10-fold cross validation with k set to 10. First, we divide all
datasets into 10 parts. Then use 9 of the parts for training, and 1 for testing.
Repeat the procedure 10 times, rotating the test datasets.

Model Evaluation

Most of the methods of machine learning evaluate the accuracy of the model
by creating a model using training datasets and applying test datasets to this
model. There are many ways to measure the quality of classification models.
Among them, we use Matthews Correlation Coefficient (Matthews 1975). It
considers true and false positives and negatives and is used as a balanced
measure which can be used although the types are of very different sizes
(Boughorbel, Jarray et al. 2017). The MCC returns a value between $-1$ and
$+1$. A coefficient of $+1$ represents a perfect prediction, $0$ no better than
random prediction and $-1$ indicates total disagreement between prediction
and observation.

From the confusion matrix, we can directly calculate the MCC:
Matthews Correlation Coefficient

\[
\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}}
\]

In this equation, TP and TN are the number of true positives and negatives, respectively. FP and FN are the number of false positives and negatives, respectively. TP means True A (targeting type)’s that are correctly classified as A. TN means all other types correctly classified as not A. FP means other types incorrectly classified as A. FN means A’s that were incorrectly classified as not. In confusion matrix, each row of the matrix represent the instance in a predicted type while each column represents the instances in an actual type (Powers 2011).
3.4 Results

Grouping

First, *GenomeTrakr Network* datasets were grouped into five pathogenic *E. coli* using Bowtie2. This process was performed in two steps with Bowtie2; 1) setting virulence factors as an index, and 2) mapping the paired-end reads with this index. We grouped the virulence factors provided by VFDB (Table 3-2) by setting index. As a result, the sizes of EAEC, EHEC, EIEC, EPEC, and ETEC groups were 34, 453, 10, 749, and 13, respectively. The number of samples of EAEC, EIEC, and ETEC was absolutely lower than that of EHEC and EPEC. So, we combined them into new type called “ETC” (Figure 3-3).

Alignment

These grouping pathogenic *Escherichia coli* were aligned with Nissle 1917 as a reference genome using SAMtools. The resulting SAM and BAM format files were converted to Bed format for a more rapid genome arithmetic (Kindlon 2017). Short readings were concatenated until there was no
alignment possible. This process corresponds to the steps (b) to (c) in Figure 3-4.
Figure 3-3. The number of five pathogenic *E. coli* samples was used in this study.
Sliding window and Transformation

The last step of data transformation consisted in Sliding Window analysis (Figure 3-4c). The size of window (with overlap in brackets) was either set to 10,000 bp (1,000 bp, 500 bp and no overlap), 5,000 bp (500 bp and no overlap) or 1,000 bp (No overlap). The divided ranges by window size were used as features for further analysis.

The size of *E. coli* whole genome is approximately 5,460,000 bp. The number of features created thus depended on the window and overlap selected sizes (as illustrated in Table 3-4). For example, a window size of 1,000 bp without overlap results in the creation of 5,460 ranges. We then performed the last preprocessing step for the analysis, consisting in coding the features as categorical variables: each window was assigned either the value ‘1’ (window is over 90% aligned) or ‘0’ otherwise (Table 3-5).
Figure 3-4. Concept of data transformation used in our analysis.

(a) It means the reference genome used for alignment. Nissle 1917 was used in this study as the *E. coli* reference genome. (b) Each one means a short read aligned by SAMtools. (c) Makes long reads by concatenating short reads. Through this process, we can find out where the alignment is and where it is not. (d) By sliding window analysis, it is converted to '1' when the sample is aligned by 90% or more and '0' when the sample is not aligned by the reference (Table 3-5).
<table>
<thead>
<tr>
<th>Window size</th>
<th>Overlap size</th>
<th># of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000 base pair</td>
<td>No overlap</td>
<td>5,460</td>
</tr>
<tr>
<td>5,000 base pair</td>
<td>500 base pair</td>
<td>10,900</td>
</tr>
<tr>
<td>5,000 base pair</td>
<td>No overlap</td>
<td>1,092</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>500 base pair</td>
<td>10,900</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>1,000 base pair</td>
<td>5,450</td>
</tr>
<tr>
<td>10,000 base pair</td>
<td>No overlap</td>
<td>546</td>
</tr>
</tbody>
</table>
Table 3-5. Example of data transformation by sliding window analysis.

<table>
<thead>
<tr>
<th>Type</th>
<th>0bp</th>
<th>1,000bp</th>
<th>2,000bp</th>
<th>3,000bp</th>
<th>4,000bp</th>
<th>5,000bp</th>
<th>6,000bp</th>
<th>7,000bp</th>
<th>8,000bp</th>
<th>9,000bp</th>
<th>10,000bp</th>
<th>11,000bp</th>
<th>12,000bp</th>
<th>1,114,000bp</th>
<th>1,115,000bp</th>
<th>1,116,000bp</th>
<th>...</th>
<th>5,448,000bp</th>
<th>5,449,000bp</th>
<th>5,450,000bp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1*</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>EHEC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EHEC</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EHEC</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EHEC</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EHEC</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>EPEC</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPEC</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPEC</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>EPEC</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>ETC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETC</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

* 1: alignment score ≥ 0.9, 0: alignment score < 0.9
Feature Selection by Chi-squared test

It is apparent from Table 3-4 that too many features have occurred indiscriminately so that removing redundant features would be desirable. Since both feature and response variables in this study are categorical, feature selection was performed using chi-square test (Kaushik 2017) as implemented in the function chi.squared from the R "FSelector package". This function was used to evaluate the importance of variables for all window sizes, one of which is shown in Table 3-6. The number of features was fixed to either 400, 200, 100, 50, 25 or 12 according to their importance, and further used for DT analysis. Raw data without feature selection was also used for analysis.
Table 3-6. Example of the importance of features

<table>
<thead>
<tr>
<th>Feature (Window position)</th>
<th>Attr_importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 ~ 10,000 bp</td>
<td>0.14448873</td>
</tr>
<tr>
<td>10,000 bp ~ 20,000 bp</td>
<td>0.36884852</td>
</tr>
<tr>
<td>20,000 bp ~ 30,000 bp</td>
<td>0.23898066</td>
</tr>
<tr>
<td>30,000 bp ~ 40,000 bp</td>
<td>0.3118214</td>
</tr>
<tr>
<td>40,000 bp ~ 50,000 bp</td>
<td>0.07125117</td>
</tr>
<tr>
<td>50,000 bp ~ 60,000 bp</td>
<td>0.60619677</td>
</tr>
<tr>
<td>70,000 bp ~ 80,000 bp</td>
<td>0.16338621</td>
</tr>
<tr>
<td>80,000 bp ~ 90,000 bp</td>
<td>0.51408799</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>5,270,000 bp ~ 5,280,000 bp</td>
<td>0.0799372</td>
</tr>
<tr>
<td>5,310,000 bp ~ 5,320,000 bp</td>
<td>0.15838282</td>
</tr>
<tr>
<td>5,390,000 bp ~ 5,400,000 bp</td>
<td>0.20305021</td>
</tr>
<tr>
<td>5,400,000 bp ~ 5,410,000 bp</td>
<td>0.04639442</td>
</tr>
</tbody>
</table>
Decision tree

We used the R packages “party” to perform classification analysis and conducted K-fold cross validation with K set to 10 for validation. The entire data was randomly divided into 10 groups. For each group, 9 subgroups were set as training datasets to learn for the classification model and 1 subgroup was set as test datasets to verify the model. This cross validation was conducted 30 times for the reproducibility of the analysis. Results were then further checked for MCC using the test datasets: after predicting the type of test set using R's predict function, the confusionMatrix function was used to evaluate the accuracy of the model. The model prediction accuracies according to R package use in function of the window size, overlap size, and the number of features are displayed in Figure 3-6. Since the number of samples was very different for each type (EHEC, EPEC and ETC), the model evaluation was conducted focusing on the MCC rather than the accuracy of the model. Unlike the 2 by 2 confusion matrix, the 3 by 3 confusion matrix is set differently for TP, TN, FP and FN according to the criteria of type (Figure 3-5).
Figure 3-5. These matrices count how many samples that truly belong to each type (columns) were predicted to belong to that type (rows) when we set the type criteria differently. (a) Criteria of type: EHEC (b) Criteria of type: EPEC (c) Criteria of type: ETC (EAEC, EIEC, ETEC)
**Figure 3-6.** The graph below is a harmonic average that calculated Matthews Correlation Coefficient (MCC) values by type (EHEC, EPEC, and ETC). If the MCC is not calculated, no value is displayed in the graph.
We obtained 3 by 3 confusion matrix by applying test datasets to the model created by training datasets. In this matrix, we calculated the MCC for each window size, overlap size, and number of features. When the type is EHEC, EPEC, ETC, the MCC is calculated and the harmonic mean of the three MCCs is used as the performance evaluation index of the model (Figure 3-6). When arithmetic average is calculated, the larger the value takes a higher weight than the smaller value. On the other hand, the harmonic average can compensate for these disadvantages and achieve fair average. The highest MCC of 0.9706 was associated with a window size of 5,000 bp (overlap size is 500 bp) and the number of features being 400. The decision tree and the confusion matrix of this result are displayed in Figure 3-7.
Figure 3-7. Decision tree model and confusion matrix when window size is 5,000 bp, overlap size is 500 bp and the number of features is 400. (a) Decision tree model using training datasets.
Figure 3-7. (b) Confusion matrix using test datasets.

Confusion Matrix and Statistics

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
<th>EHEC</th>
<th>EPEC</th>
<th>ETC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EHEC</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>EPEC</td>
<td>2</td>
<td>72</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>ETC</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Overall Statistics

Accuracy : 0.9811  
95% CI : (0.9335, 0.9977)  
No Information Rate : 0.6792  
P-Value [Acc > NIR] : 2.023e-15  
Kappa : 0.9566  
Mcnemar's Test P-Value : NA

Statistics by Class:

<table>
<thead>
<tr>
<th></th>
<th>Class: EHEC</th>
<th>Class: EPEC</th>
<th>Class: ETC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.9394</td>
<td>1.0000</td>
<td>1.000000</td>
</tr>
<tr>
<td>Specificity</td>
<td>1.0000</td>
<td>0.9412</td>
<td>1.000000</td>
</tr>
<tr>
<td>Pos Pred Value</td>
<td>1.0000</td>
<td>0.9730</td>
<td>1.000000</td>
</tr>
<tr>
<td>Neg Pred Value</td>
<td>0.9733</td>
<td>1.0000</td>
<td>1.000000</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.3113</td>
<td>0.6792</td>
<td>0.009434</td>
</tr>
<tr>
<td>Detection Rate</td>
<td>0.2925</td>
<td>0.6792</td>
<td>0.009434</td>
</tr>
<tr>
<td>Detection Prevalence</td>
<td>0.2925</td>
<td>0.6981</td>
<td>0.009434</td>
</tr>
<tr>
<td>Balanced Accuracy</td>
<td>0.9697</td>
<td>0.9706</td>
<td>1.000000</td>
</tr>
</tbody>
</table>
Genomic Regions

The highest harmonic average of each of the three types (EHEC, EPEC and ETC) is 0.9706 when the window size is 5,000 bp, the overlap size is 500 bp and the number of features is 400. We propose genomic regions focusing on decision tree model (Figure 3-7a) and 3 by 3 confusion matrix (Figure 3-7b) under these conditions. The nodes in this model represent genomic regions. In the model, ‘1’ indicates that the alignment score is more than 0.9 and ‘0’ is less than 0.9 relative to the reference genome, Nissle 1917. For example, the root node “S3358500E3363500” corresponding to the genomic region starting at 3,358,500 bp and ending at 3,363,500 bp. The further nodes and their corresponding genomic regions starting from this region further participate in the classification of the *E. coli* in Figure 3-7a. The method of interpreting nodes of this model as genomic regions is as follows; Compute the alignment score with Nissle 1917 for an unknown *E. coli* genome. If the unknown genome is aligned at 90% or more from 3,368,500 bp to 3,358,500 bp (S3358500E3363500) and 90% or more from 1,748,500 bp to 1,743,500 bp (S1743500E1748500), it is classified as EHEC. Here we suggest genomic regions that can be classified as EHEC by 3,358,500 bp to 3,363,500 bp and
1,743,500 bp to 1,748,500 bp. When unknown genome is aligned 90% or more from 3,358,500 bp to 3,363,500 bp and less than 90% from 1,743,500 bp to 1,748,500 bp, we can classify this genome as EPEC. We conclude that there is a genomic region from 3,358,500 bp to 3,363,500 bp and no genomic region from 1,743,500 bp to 1,748,500 bp in order to classify the unknown genome as EPEC.

As a result, these nodes (genomic regions) in the tree model constitute a novel way to classify pathogenic E. coli. Using DT analysis, it has been determined that these genomic regions play an important role in classifying pathogenic E. coli and it would thus be practicable to use these regions as primers.

3.5 Discussion

E. coli is commonly found in the lower intestine of warm-blooded animals, including humans (Marler 2008). While most E. coli strains are harmless, some can cause serious food poisoning. Probiotics E. coli strain Nissle 1917 was used as a reference genome among the non-disease-causing strains.
Before the analysis, we thought that a smaller window size gave better classification performances. However, smaller window sizes were associated with more features, this latter fact having adverse effects in building a robust classification model. Feature selection made a substantial contribution in improving the prediction performance of the model. Reducing the window size and performing feature selection at the same time resulted in a classification model with better classification prediction performance.

In this paper, we propose a new method for classifying pathogenic *E. coli*. First, to evaluate the performances of the decision tree models the datasets were divided into train and test datasets. We conducted 10-fold cross validation. The decision tree model was built with the train datasets and test datasets were then applied to determine how accurately the results were predicted. The performances of the decision tree models (fitness of the model and classification prediction performances) obtained “party” packages in R.

This study has several limitations. First, the size of the datasets issued from different types of *E. coli* is not evenly distributed. Compared to EHEC and EPEC, the number of data in EAEC, EIEC and ETEC types is clearly insufficient (as an extreme example EIEC sample size is more than 70 times
smaller than EPEC sample size) which can lead to overfitting or underfitting
the model (Brownlee 2016). As the model was calibrated using these highly-
unbalanced datasets the classification result might have been biased in favor
of EHEC and EPEC. In other terms, data belonging to the least represented
groups during data training (EAEC, EIEC, and ETEC) are more susceptible
to be mistakenly affected to the more represented EHEC and EPEC groups.
As a result, this uneven distribution of datasets can increase false positive rate
(Type I error). To create a more accurate prediction model, the five types of
*E. coli* data should be provided evenly. To overcome this shortcoming, we
have created the model that focuses on classifying EHEC and EPEC, which
have many samples. So we combined EAEC, EIEC, and ETEC, which have
small samples, to create a new type of ETC. ETC types were used for this
analysis in place of the three types (EAEC, EIEAC, and ETEC).

The second limitation is ‘NaN’. NaN, standing for ‘Not A Number’, is a
numeric value representing an undefined or un-representable value. In this
analysis, 0/0 is undefined as a real number, and so represented by NaN. This
NaN occurred in some metrics (Precision, Accuracy, Recall, etc.) computed
from the confusion matrix for validation. When the model validation was
performed, the test datasets were selected at random. In this case, it happened accidentally that ‘0’ of a certain type (mainly ‘ETC’) of reference was selected in the test datasets and the model predicted ‘0’ for this type. When NaN occurred, the value of MCC, which is an index for evaluating the performance of the model, cannot be obtained. As a result, we cannot evaluate the performance of the model. For this, MCC was calculated except for the mode when NaN occurred in the confusion matrix. If the MCC didn’t exist in EHEC, EPEC and ETC (like ‘NaN’), it was not shown in Figure 3-6.

Without the experimental method, we classified the pathogenic E. coli in-silico only based on NGS data. This study thus proposes a novel method in classifying pathogenic E. coli as a new approach and will hopefully contribute in helping in the treatment of pathogenic E. coli-related diseases.

We herein propose new criteria for the classification of pathogenic E. coli (EAEC, EHEC, EIEC, EPEC and ETEC). Since the number of samples was insufficient, we created a new type, ETC, which combined EAEC, EIEC and ETEC. We conducted a decision tree analysis, a form of classification analysis, with 3 types of pathogenic E. coli data. To determine the exact performance of the model, we calculated the MCC from the confusion matrix.
as well as the classification accuracy. The resulting classification model exhibited a high accuracy of 98.11% and MCC of 0.9706 and genomic regions classifying pathogenic *E. coli* were identified. Further research is required to identify the functions of these regions so that they may be used as markers or primers.
Reference


Luo, W., et al. (2012). "Genome-wide association analysis of meat quality


estimation and model selection. Ijcai, Montreal, Canada.


Versalovic, J. and J. R. Lupski (2002). "Molecular detection and


국문초록

유전체 다형성을 이용한 연관과 분류를 위한 선형 및 비선형 모델의 적용

장동성

협동과정 농생명유전체학전공

서울대학교 대학원 농업생명과학대학

나는 유전적 다형성으로부터 유전적 변이들을 발굴하기 위해 연구를 수행했다. 첫번째 연구에서 SNP 마커들과 표현 형질 간의 연관성을 찾기 위해 전장 유전체 연관 분석(Genome-Wide Association Studies, GWAS)을 실시했다. 이 과정을 위해 PLINK 프로그램을 이용해서 선형 모델을 만들었다. 두번째 연구에서는 종을 분류할 수 있는 유전적 변이를 발굴하기 위해 유전적 영역을 이용한 분류를 위한 의사결정 나무를 실시했다. 의사결정 나무는 비선형모형으로, 특히 다분화(multiclass) 종속 변수에 대해서 적용 가능한 장점이 있다 (Kim 2003; Ozaki 2014). 이번 연구에서 종속 변수는 병원성 대장균인데 증상에 따라서 다섯 가지로
다분화 되어 있기 때문에 의사결정 나무를 적용해 분류 분석을 실시했다.

즉, 단일 염기 다형성 (SNP) 수준에서 두 가지 표현형을 사용하여 다변량 GWAS를 수행하여 돼지 전장 유전체로부터 변이를 검출하는 연구를 실시했다. 그리고 유전체 영역 수준에서는 의사결정나무 분석을 수행하여 대장균 전장 유전체의 유전적 영역을 이용하여 이들 병원성 대장균의 종을 분류 할 수 있는 유전적 변이를 발굴하기 위한 연구를 수행했다.

2장에서는 버크셔 (Berkshire) 돼지와 듀록 (Duroc) 돼지의 SNP 마커들과 표현 형질을 이용하여 다변량 GWAS를 수행했다. 일반적으로 버크셔는 듀록 보다 부드럽다고 한다. 나는 어떤 유전적 변이가 이 두 품종의 돼지 고기에서 변이를 유발하는지 알아내기 위해 연구를 수행했다. 이번 연구에서 목표 표현 형질은 고기의 부드러움이다. 이 부드러움은 다양한 형질이 관여하는 복잡한 특성이기에 일반적으로 사용되는 일변량 GWAS가 아닌 다변량 GWAS를 수행했다. 다변량 GWAS의 결과를 토대로 부드러움과 관련된 SNP 마커 (ASGA0033314)를 발굴 했고, 이 마커와 관련된 후보 유전자들 (MEP1A, ARPC1A, ENAH, APC2)을 제시했다. 이 연구에서 찾은 부드러움과 관련된 유전자들은 돼지 고기의 육질을 개선하는 데 도움이 될 것이다.
3 장에서 병원성 대장균을 분류하는 분석을 수행했다. 병원성 대장균 중에서도 식중독을 일으킨다고 알려진 설사성 질환과 관련 있는 대장균 5 가지를 이번 연구에서 사용했다. 일반적으로 병원성 대장균은 독성 인자의 유무로 분류된다. 이번 연구는 새로운 접근법으로서 기존 방법과 다른 분류 방법을 제안하기 위해 수행되었다. 나는 병원성 대장균을 분류할 수 있는 유전체 영역을 얻기 위해 의사결정나무 분석을 실시했다. 분석 결과 얻은 유전체 영역을 대장균을 분류할 수 있는 새로운 방법으로 제시했다. 이 영역들은 아직 기능이 밝혀지지 않은 부분들이 다수 존재하지만, 훗날 기능이 밝혀진다면 마커나 프라이머로써 사용이 가능할 것으로 기대된다. 이 연구는 슬라이딩 윈도우 방법으로 병원성 대장균을 분류하는데 의의를 두고 병원성 대장균 관련 증상을 치료하는데 도움을 줄 것으로 기대한다.

주요어: 유전적 다형성, 다변량 전장 유전체 연관 분석, 분류, 유전체 영역, 슬라이딩 윈도우, 의사결정 나무

학번: 2016-24132