이 학 석 사 학 위 논 문

Application of statistical analysis in transcriptomic and metagenomic data

전사체와 메타지놈 데이터 분석을 위한
통계적 방법의 활용

2019 년 8 월

서울대학교 대학원
협동과정 생물정보학과
방 소 현

Application of statistical analysis in transcriptomic and metagenomic data

By

Sohyun Bang

Supervisor: Professor Hee-bal Kim

Aug, 2019

Interdisciplinary Program in Bioinformatics

Seoul National University

전사체와 메타지놈 데이터분석을 위한 통계적 방법의 활용

지도교수 김 희 발

이 논문을 이학석사 학위논문으로 제출함

2019 년 7 월

서울대학교 대학원

협동과정 생물정보학과

방 소 현

방소현의 이학석사 학위논문을 인준함

2019 년 7 월

위 원 장 이 용 환 (인)

부위원장 김 희 발 (인)

위 원 조 서 애 (인)

Abstract

Application of statistical analysis in transcriptomic and metagenomic data

Sohyun Bang

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

With the advance in sequencing technology, genomics, transcriptomics, proteomics, epigenomics and metagenomics study genetic materials on the genome-wide scale. Transcriptome and metagenomics have common tools and methods for analysis because they use quantitative data. Analysis of transcriptome data aims at quantifying gene expression and finding differentially expressed genes (DEG) under the certain condition. To detect DEG and trait-associated genes, various statistical methods and tools have been developed and some of them are widely being used. As analysis of metagenome data also use quantified abundance of microorganisms, some developed tools for transcriptome analysis were also applied in metagenome data. In this thesis, I described how the statistical methods were employed to solve biological problems in quantitative data.

Analysis of quantitative data recently employed machine learning based methods to predict traits including disease and healthy status. Especially, as gut

microbiome is associated with host's health, several studies suggested the possibility to diagnosis the diseases using abundance and kinds of microorganisms living in gut. In Chapter 2, machine learning based multi-classifier algorithms were established and evaluated to classify several diseases using gut microbiome data. LogitBoost algorithms and using abundance of microorganisms at genus level showed the highest performance. By selecting microorganisms to enhance the performance, the selected microorganisms were suggested as markers to classify various disease simultaneously.

Gut microbiome is used as significant marker not only in human health but also in domestic animals. For example, microbial communities altered by feeds in broiler chickens. In Chapter 3, the effect of A. hookeri on gut microbiome in young broiler chickens was investigated. Statistical test revealed that the composition of microbiome was altered by supplement with A. hookeri leaf. The modulated gut microbiome by leaf was correlated with growth traits including body weight, bone strength, and infectious bursal disease antibody.

For more accurate analysis of quantitative data, accurate quantification of genetic materials is essential. Chapter 3 suggests the cause of mis-quantification of mRNAs and solutions to reduce the mis-quantified expression. Long non-coding RNAs, which are overlapped with mRNAs in genomic position, can be mis-quantified to the overlapped mRNAs. Simulation showed the degree of errors by such mis-quantification. Tools for alignment and quantification were compared to reduce the error and achieve more accurate quantification for transcriptome.

Key words: Gut microbiome, Machine learning, Disease classification, Growth performance, RNA-Seq, long non-coding RNAs

Student number: 2016-20462

Contents

List of Tables

List of Figures

Chapter 1. Literature Review

1.1 Machine learning approaches for gut microbiome data

1.1.1 Gut microbiome related to diseases

Gut microbial community is associated with host digestion, nutrition and even regulation of host immune system (Rooks and Garrett 2016). Microbiota facilitates the development and function of the immune cells at both mucosal and nonmucosal sites (Belkaid and Hand 2014). This affects immune system through whole body as well as gut immune system (Belkaid and Hand 2014). The intestinal microbiota adheres to the mucosal layer on the inner wall of the intestinal tract to form a barrier (Ferrario, Taverniti et al. 2014). This helps to repress the growth of pathogenic bacteria and prevent pathogens from entering the body(Nagai, Morotomi et al. 2009). The disruption of intestinal homeostasis will lead to disease, which is caused by antibiotics, diet and other factors (Eeckhaut, Machiels et al. 2012). Previous studies have shown that the gut microbiota is closely associated with obesity, type 2 diabetes, liver diseases (Biddle, Stewart et al. 2013).

1.1.2 Feature selection

Feature selection is a procedure that remove redundant features from total features to retrieve a subset of the input set (Jain and Zongker 1997). The main objective of feature selection is improving the prediction or classification accuracy. In the context of classification, there are several categories of feature selection techniques (Kuo 2013). First, filter method extract the relevant features by removing the least interesting features. It can also make the model faster and more cost-effective, and provide an deeper understanding into the

underlying process that generated the input data (Guyon and Elisseeff 2003, Kuo 2013). This method is particularly effective in computation time and independent of the classifier. However, it does not consider the interaction between feature and classifier (Yu and Liu 2003).

Next, wrapper method detects the possible interactions between features. It searches an optimal subset from all feature subsets by assessing the classification performance using possible feature subsets. One of Advantages for wrapper approaches is dependency to classification model. On the other hand, disadvantage of wrapper method is that it has a higher risk of overfitting than filter method and are very computationally intensive to compute higher number of feature subsets (Das 2001).

1.1.3 Multiclass classifiers

KNN, LogitBoost, LMT and SVMs with sequential minimal optimization (SMO) have previously shown high multi-group classification performance were employed including (Hsu and Lin 2002, Landwehr, Hall et al. 2005). The KNN implies a classifier capable of multi-groups classification. The LogitBoost is a developed boosting algorithm that can handle multiclass problems by considering multiclass logistic loss(Friedman, Hastie et al. 2000). The LogitBoost has been applied to predict protein structural classes(Cai, Feng et al. 2006) and places of origin for pigs with high performance(Kim, Seo et al. 2015). The LMT is based on a regression tree that has logistic models on the leaves(Landwehr, Hall et al. 2005). In predictions related to medical application including prediction of response to antiretroviral combination therapy or autism spectrum disorder, LMT showed an advantage over the other methods

(Altmann, Beerenwinkel et al. 2007, Jiao, Chen et al. 2010). The SMO has been shown to be an effective method for SVM on classification tasks without a quadratic programming solver. The KNN and SVM classifiers are the most widely used methods and they have been applied successfully in numerous studies(Liu, Hsiao et al. 2011, Kim, Seo et al. 2015).

1.2 Community analysis in the metagenomic data

1.2.1 Analysis for microbial communities

The metagenomics is an approach for simultaneous identification of microbiomes with cultivation-independent assessment. Because of the limitation of the sequencing length, some variable regions of 16s rRNA were used as phylogenetic marker to distinguish taxon(Poretsky, Rodriguez et al. 2014). Taxonomic analysis for bacteria is regularly performed using 16S data derived from varying sequencing technologies (ie, 454 pyrosequencing as well as Illumina, Solid and Ion Torrent). Commonly used tools for 16S data analysis and denoising include QIIME, Mothur, and MEGAN (Huson, Auch et al. 2007, Schloss, Westcott et al. 2009, Caporaso, Kuczynski et al. 2010). They generates overlapped fragments by assembly among paired-end reads, filters bad quality overlapped sequences, and assigns the overlapped sequence to phylogenetic groups, and measures the abundances in each OTUs.

Using taxonomy assigned bacteria, detecting differentially abundant microbiome similar to DEG detection can be performed (White, Nagarajan et al. 2009). Similar statistical methods with RNA-Seq can be utilized in

metagenome. Furthermore, several statistics for summarization of microbial communities including α-diversity and β-diversity can be used (Lozupone and Knight 2005). Correlation analysis between the abundance of bacteria and traits can be performed (Zeng, Han et al. 2015).

1.3 Quantification of mRNAs and lncRNAs

1.3.1 RNA sequencing and its application

Since 1970, gene expression profiling has been actively performed on a variety of platforms. The biological techniques for quantifying gene expression have undergone several major revolutions, and sequencing-based methods have been used predominantly in the past decade. Especially, the gene expression levels give one of the more direct quantitative inference of the relationship between phenotype and biological markers. It is currently considered as the most practical and representative biological information (Schena, Shalon et al. 1995, Wang, Gerstein et al. 2009). The advance in sequencing technology such as CAGE and RNA-seq enabled genome-scale investigation of the genes (Kawaji, Lizio et al. 2014). In RNA-seq platform, reads produced by transcripts were counted and processed by normalization and testing. Normalization adjusts systematic biases to determine relative expression (Robinson and Oshlack 2010). TMM normalization is most widely used recently (Robinson and Oshlack 2010). Calculated normalized values are used in generalized linear model to detect DEGs. For this, EdgeR and DESEQ2 have been developed and used (Robinson, McCarthy et al. 2010, Love, Huber et al. 2014).

1.3.2 Characteristics of lncRNA

The lncRNAs, which are non-protein-coding RNAs over 200bp in length, have been recently discovered, and are expected to be widely spread across the genome (Mercer, Dinger et al. 2009, Kung, Colognori et al. 2013). Some reports have elucidated the importance of their functional regulation of transcription and splicing (Mercer, Dinger et al. 2009, Guil and Esteller 2012). For instance, they can serve as competing endogenous RNAs to modulate the concentration and biological function of mRNAs and miRNAs. Additionally, the lncRNAs can also directly bind to target genes or act as scaffolds for transcription factors and histone modifiers to activate/inhibit the expression of target genes.

The lncRNAs are expected to be capped, polyadenylated, and widespread across the genome (Yang, Wen et al. 2015). They are frequently involved in regulatory functions of transcriptional, post-transcriptional and epigenetic processes (Guttman, Amit et al. 2009, Chen and Carmichael 2010). There are mainly four ways that all lncRNAs execute their regulatory functions: as decoys, guides, scaffolds, or signals (Wang and Chang 2011). For example, during the early stages of embryonic development in mammals, the X chromosome in the female are inactivated to achieve the same expression levels of X-chromosomal genes in male mammals. This phenomenon is called the X-chromosome inactivation (XCI), and the lncRNA Xist (X-inactive specific transcript) is reported to play an essential role in the initiation of XCI (Lv, Yuan et al. 2016). The lncRNAs also mimic receptor or endogenous target sites in both animals and plants to regulate structural transformation and miRNA expressions (Franco-Zorrilla, Valli et al. 2007, Kino, Hurt et al. 2010, Heo, Lee et al. 2013). As illustrated, the involvement of lncRNAs are being elucidated at both

structural and functional levels, but we are far from understanding all the underlying biological mechanisms and processes in relation to lncRNAs.

1.3.3 Overlap problems in quantifying transcripts

Comprehensive analyses of ambiguous read problems such as isoform problem (Li, Ruotti et al. 2010) and mRNA-mRNA overlap problem has been attested, and in vitro or in silico solutions are proposed to disburden those problems (Sun, Yang et al. 2015). Under in vitro conditions, a strand-specific protocol during cDNA synthesis is a partial solution for the overlap problems in the complementary strand (Sigurgeirsson, Emanuelsson et al. 2014). Also, there are methods such as re-designing the microarray probes, to consider the expression levels in the overlap region (Yelin, Dahary et al. 2003). On the other hand, under in silico conditions, overlapped region problems are handled through algorithms in the quantification steps that consider repetitive features induced by alternative splicing and overlap region (Sun, Yang et al. 2015, Schuierer and Roma 2016). Although there were some attempts on mitigating the overlap problem, a systematic investigation of such errors caused by mRNA and lncRNA overlap needs to be conducted in several species for developing an improved version from the current quantification algorithm.

This chapter was published in *Scientific Reports*
as a partial fulfillment of Sohyun Bang's M.Sc program.

Chapter 2. Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data

2.1 Abstract

Diseases prediction has been performed by machine learning approaches with various biological data. One of the representative data is the gut microbial community, which interacts with the host's immune system. The abundance of a few microorganisms has been used as markers to predict diverse diseases. In this study, we hypothesized that multi-classification using machine learning approach can distinguish the gut microbiome from following six diseases: multiple sclerosis, juvenile idiopathic arthritis, myalgic encephalomyelitis/chronic fatigue syndrome, acquired immune deficiency syndrome, stroke and colorectal cancer. We used the abundance of microorganisms at five taxonomy levels as features in 696 samples collected from different studies to establish the best prediction model. We built classification models based on four multi-class classifiers and two feature selection methods including forward selection and backward elimination. As a result, we found that the performance of classification is improved as we use the lower taxonomy levels of features; the highest performance was observed at the genus level. Among four classifiers, LogitBoost–based prediction model outperformed other classifiers. Also, we suggested the optimal feature subsets at the genus-level obtained by backward elimination. We believe the selected feature subsets could be used as markers to distinguish various diseases simultaneously. The finding in this study suggests the potential use of selected features for the diagnosis of several diseases.

2.2 Introduction

Machine learning technology has been applied in various fields and has become a useful strategy in the field of biotechnology, especially for predicting diseases and supporting medical diagnosis(Kukar, Kononenko et al. 1999, Cruz and Wishart 2006, Sajda 2006). In order to predict diseases, biological data including gene expression, genotype, and methylation level can be employed(Cho and Won 2003, Knights, Costello et al. 2011). Moreover, the realms of biological data have been extended to include the microbial communities due to their association with the host's immune system(Rooks and Garrett 2016). Microbial communities facilitate the development and function of the immune cells at both the mucosal and nonmucosal sites(Maranduba, De Castro et al. 2015). Their regulation of the immune system is involved in various diseases(Kinross, Darzi et al. 2011). Such association have been identified in diseases like multiple sclerosis (MS), juvenile idiopathic arthritis (JIA), myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), stroke, acquired immune deficiency syndrome (AIDS), and colorectal cancer (CRC)(Baxter, Koumpouras et al. 2016, Baxter, Ruffin et al. 2016, Di Paola, Cavalieri et al. 2016, Giloteaux, Goodrich et al. 2016, Jangi, Gandhi et al. 2016, Noguera-Julian, Rocafort et al. 2016).

Some of the well-known researches have attempted to establish a disease-prediction model based on the gut microbiome data from healthy individuals and patients, and have discovered that gut microbiome data can be applied to

predict specific diseases(Giloteaux, Goodrich et al. 2016). Patients with irritable bowel syndrome and healthy individuals were classified using Random forest algorithm(David, Maurice et al. 2014). Other diseases such as liver cirrhosis, colorectal cancer, inflammatory bowel diseases, obesity, and type 2 diabetes were distinguished with a healthy status using machine learning approaches(Pasolli, Truong et al. 2016). Most of these studies have focused mainly on diagnosing only one disease, and so far, there have been few attempts to predict multiple diseases at once.

The potential of multi-classification using microbiome data is being shown in recent studies(Statnikov, Aliferis et al. 2005, Liu, Hsiao et al. 2011). In the case of classifying various body parts, a previous study performed multi-classification based on KNN and probabilistic neural networks(Statnikov, Aliferis et al. 2005). In another study, multi-classification of three different diseases was demonstrated using selected metagenomic biomarkers(Singh, Chang et al. 2017). Similarly, in our study, we hypothesized that various diseases can be classified using gut microbiome data from 16S rRNA sequencing. To investigate the possibility of classification on various diseases based on the microbial community, we collected a total of 1,079 metagenome data from healthy individuals and patients with following diseases in six studies: MS, JIA, ME/CFS, AIDS, CRC, and Stroke. To combine data, we preprocessed data using normalization and statistical method. We classified six diseases listed above using the abundance of microorganisms at the phylum, class, order, family, and genus levels as features. We built classification models based on the multi-class classifiers such as LogitBoost, support vector machine (SVM), K nearest neighbor (KNN) and logistic model tree (LMT). Moreover, we

constructed a feature subset using two feature selection methods. We compared the performance of classification in three factors: 1) taxonomy levels of features, 2) four classifiers and 3) feature selection methods.

2.3 Materials and Methods

2.3.1 Collection of the gut microbiome data related to six diseases

For disease prediction based on the metagenome data sets of gut microbial communities, large numbers of metagenome samples were collected from the European Bioinformatics Institute (EBI) database (https://www.ebi.ac.uk/metagenomics/). To minimize the biases caused by different experimental protocols, data were collected with several criteria: (1) 16S rRNA based metagenome data through the stool sampling, which is widely used approach at present, (2) sequencing platforms including 454 and Illumina's, (3) using first measurement in case of longitudinal data to ensure independence assumption and (4) EBI pipeline v2.0 or v3.0 (https://www.ebi.ac.uk/metagenomics/pipelines/3.0) for identifying and quantifying the OTUs. In EBI pipeline, several tools used are as following: (1) Trimmomatic (v0.32)(Bolger, Lohse et al. 2014) for quality check and trimming of low quality reads; (2) SeqPrep (v1.1)(Hunter, Corbett et al. 2014) to merge paired-end reads to generate overlapped read; (3) rRNASelector (v1.0.1)(Lee, Yi et al. 2011) to filter out of non-ribosomal RNA; (4) QIIME(v1.9.0)(Caporaso, Kuczynski et al. 2010) for OTU identification and quantification. From this pipeline, gut microbial communities data was generated at various taxonomic levels such as phylum, class, order, family, and genus based on the Greengenes 16S rRNA database(DeSantis, Hugenholtz et al. 2006).

2.3.2 Preprocessing of the metagenomic data derived from different studies

Samples with less than 5% of the average number of reads were removed. The abundance of microorganisms at five taxonomy levels including phylum, class, order, family, and genus levels was used as features. We performed a TMM normalization for the abundance of features using edgeR(Robinson, McCarthy et al. 2010). To reduce heterogeneity across different studies, the features showing differential abundance of healthy samples between six studies were removed. We performed a log- likelihood ratio test by considering the abundance of features as negative binomial distribution(Heo, Seo et al. 2016). In the statistical test, FDR approach was used to adjust multiple testing error(Benjamini and Hochberg 1995) and 5% significance level was used for a significant result.

We further normalized the abundance with quantile normalization to produce a similar distribution of samples(Bolstad, Irizarry et al. 2003). For quantile normalization, two types of baselines can be considered to calculate normalized values: (1) global mean vector derived from each quantile of features and (2) specific baseline vector. As we assumed that distribution of all control samples are similar, the second approach was employed using only the healthy samples to create the baseline(Wu and Aryee 2010).

2.3.3 Classifiers to distinguish various diseases using the gut microbial data

We performed classification analysis with the four classifiers, implemented in the RWeka package of the R software(Hornik, Zeileis et al. 2007) with the command line of "IBk(class~.,data= InputData, control = Weka_control(K =Selected Parameter), na.action=NULL)",

“LogitBoost(class~.,data= InputData, control = Weka_control(I = Selected Parameter), na.action=NULL)”, “LMT(class~.,data= InputData, na.action=NULL)”, and SMO(class~.,data=InputData, control = Weka_control(K = list(kernel, G = Selected Parameter), C = Selected Parameter), na.action=NULL) for KNN, LogitBoost, LMT, and SVM. To assess the performance of classification, 10-fold cross-validation was used.

To select a parameter for the classifier, we used a greedy method that explores all parameter and used the parameter with the best performance. In KNN, parameter K was chosen in {3, 5, 7, 9, 11, 13, 15} (Supplementary Table 2.1). In LogitBoost, the parameter I was selected in the range from 1 to 40 (Supplementary Table 2.2). In SVM (for RBF kernel), the parameter G and parameter C were regulated in {1e-4, 1e-3…, 10} and {0.1,1,…,1000} respectively (Supplementary Table 2.3). The parameters with the highest accuracy were chosen for each taxonomy level (Supplementary Table 2.4). For the parameters with same accuracy, the one with lower value was selected.

2.3.4 Feature selection using wrapper method

We searched for a feature subset that enhances performance of classification through a wrapper feature-selection approach(Kuo 2013) including FS and BE(Kim, Seo et al. 2015). In FS, starting from the single feature with the highest accuracy, we added the feature that improves the performance the most. We continued to add features one-by-one until no more feature is left to be added. In BE, starting with all features we subtracted features one-by-one to give the highest accuracy. With the feature selection process, we obtained the feature subset showing the highest accuracy.

2.4 Results

2.4.1 Preprocessing of data to reduce biases from meta-analysis

Metagenome data from 1,079 individuals were collected for the healthy (control samples) and patients with one of six diseases including MS, JIA, ME/CFS, AIDS, Stroke and CRC (Table 2.1). The study for HIV produced the highest number of average reads (89.9M) while the study for Stroke had the lowest (4.9M). Out of all individuals, six individuals with less than 7067.68 reads (< 5% of the average) were removed. Thus, the total of 1,073 individuals-696 patients and 377 healthy samples-was used for the further analysis. The abundance of microorganisms at the phylum, class, order, family, and genus levels for 1,073 samples were normalized to correct for variations arising from use of different studies (Figure 2.1A). After Trimmed Mean of M values (TMM) normalization for the abundance of microorganisms, we compared the abundance of healthy samples from six studies. For the reason to minimize the study-dependent differences, we removed the microorganisms that are differentially abundant between studies (false discovery rate (FDR) < 0.05). Average of 16% of bacteria (5, 21, 42, 74 and 199 at the phylum, class, order, family, and genus levels, respectively) were remained (Figure 2.1 B). To further normalize the microbiome abundance of samples from different studies, quantile normalization was performed using the healthy samples as the baseline. The normalized abundance of microorganisms for 696 samples obtained in this preprocessing step was considered as features in the subsequent classification analysis.

Table 2.1 Summary of collected metagenome studies

SRA_study	Disease	Body site	# of case samples	# of control samples	Average reads per sample (std)
ERP010458	Stroke	Gut	141	92	4.9M(0.4M)
ERP013262	JIA	Gut	29	29	9.2M(2M)
ERP014628	ME/CFS	Gut	49	39	52.5M(17.1M)
SRP068240	HIV1	Gut	191	33	89.9M(69.9M)
SRP073172	CRC	Gut	263	141	14.2M(10.3M)
SRP075039	MS	Gut	29	44	31.2M(5.5M)

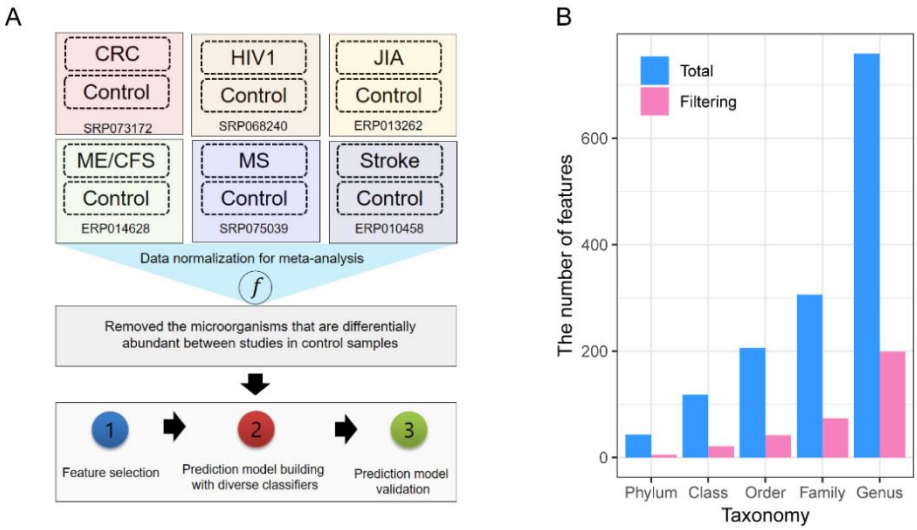


Figure 2.1. Experimental design and data processing for meta-analysis

(a) A diagram representing a whole experimental design for this research. This research consists of two major steps for analysis: (1) The process of normalization and removing features for meta-analysis; (2) The step of classification analysis to predict six diseases in integrated metagenome data across the six diseases. **(b)** Number of features at five taxonomy levels. "Total" represents the total number of features before preprocessing of data. "Filtering" represents the number of features after steps for removing features in preprocessing of data.

2.4.2 Classification performance at five taxonomy levels

To elucidate the effect of different taxonomy levels on on the classification, we assessed the performance of the classification using different sets of features such as the abundance of microorganisms at the phylum, class, order, family, and genus levels. The average of accuracies of four classifiers including KNN, LMT, LogitBoost and SVM was improved as we used the lower taxonomy levels as features (Figure 2.2A). The average of accuracies at the phylum, class, order, family and genus levels were 55, 69.9, 76.5, 80.4 and 90.4 % respectively. The accuracy at the genus level was 35.4% higher than that at the phylum level. On the other hand, the difference of accuracies between classifiers with highest accuracy (LogitBoost) and lowest accuracy (KNN) was 11.92 %. Thus, we found out that the effect of taxonomy levels on the classifier performance was greater than that of using different classifiers.

We assumed that some of the microorganisms used in the above classification might not be associated with the diseases because only a few microorganisms were found to be closely related to human health or disease(Carbonero, Benefiel et al. 2012). Hence, we performed feature selection to find features that can classify diseases more accurately. For feature selection, we used forward selection (FS) and backward elimination (BE) in four classifiers with microbial abundance at five taxonomy levels. Feature selection enhanced accuracies by 2.6%, 2.4% and 2.7% at the order, family and genus levels, respectively, while its effects were not as remarkable in phylum and class levels (0.6% and 0.4% enhanced) (Figure 2.2B). The highest accuracy improvement of 2.7% due to feature selection was observed when using features of abundance at genus level. By feature selection, 5, 21, 42, 74, and 199 number of features were reduced to 2.75, 16.5, 29.1, 45.3, and 139.5 on

average in phylum, class, order, family and genus levels, respectively (Figure 2.2C). The highest number of features was removed at the genus level. Considering the increase of accuracies and number of reduced features, feature selection was more effectively performed at the genus level.

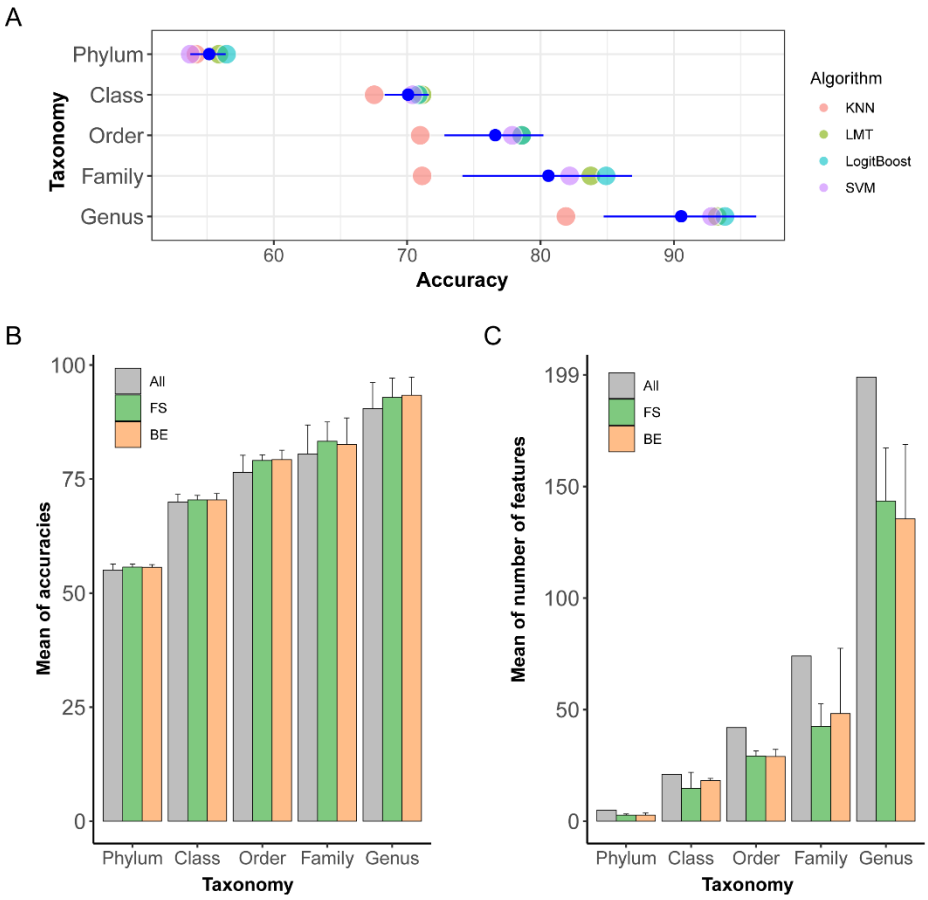


Figure 2.2 Classification performance by taxonomy levels and feature selection methods.

(a) Accuracies by taxonomy levels. Individual dots symbolize the accuracy of four classifiers. Blue dots with error bar represents the mean of the accuracies in each taxonomy. **(b)** Mean of Accuracies in four classifiers by taxonomy levels and feature selection method. The color of bars shows the feature selection method. "All" indicates that all features without feature selection are used for classifications. "FS" and "BE" indicates the features subset from FS and BE respectively. Error bar represents the standard error of accuracies at each taxonomy level and feature selection method. **(c)** Mean of number of features in four classifiers by taxonomy levels and feature selection method.

2.4.3 Comparison of classification performance at the genus level

We compared classifiers and feature selection methods based on the performance at the genus level which showed the highest performance among five taxonomy levels. The classification was conducted using 10-fold cross-validation (CV), and accuracies were averaged over three runs of 10-fold CV. Four classifiers affected the performance of classification (Figure 2.3A). The average of accuracy was the highest in LogitBoost (93.6%) followed by LMT (92.4%), SVM (91.6%), and KNN (81.5%). The difference of accuracies between classifiers with the highest accuracy (LogitBoost) and that with the lowest (KNN) was 12%. In Figure 2.2A, the difference in performance between LogitBoost and KNN increases as the taxonomy level gets lower. Regarding this aspect, the large difference (12%) between LogitBoost and KNN might come from the highest feature number at the genus level.

When we use the optimal feature sets from FS and BE, the average accuracies of the four classifiers were increased from 90.4% to 92.9% and 93.3% (FS and BE). Especially, the accuracies from KNN algorithms showed a remarkable increase from 81.8% to 86.7% and 87.5% when FS and BE were used. In all four classifiers, BE enhanced higher accuracies than FS by 0.09%, 1.19%, 0.09% and 0.43% in LogitBoost, LMT, SVM, and KNN, respectively. In LMT classifier, BE achieved the most effectively enhanced accuracies. The average number of features was reduced from 199 to 143.5 and 135.5 (FS and BE, respectively) across four classifiers (Figure 2.3B). Even though BE decreased the number of features much more compared to FS on average, the reduced number of features did not follow this trend in all classifiers. FS effectively reduced the number of features in LogitBoost algorithms, while BE did in LMT algorithm. In summary, performing feature selection enabled us to

obtain the subset of features which enhanced the overall performance of the classification in all classifiers. More importantly, higher accuracy was achieved when a lower number of features were used.

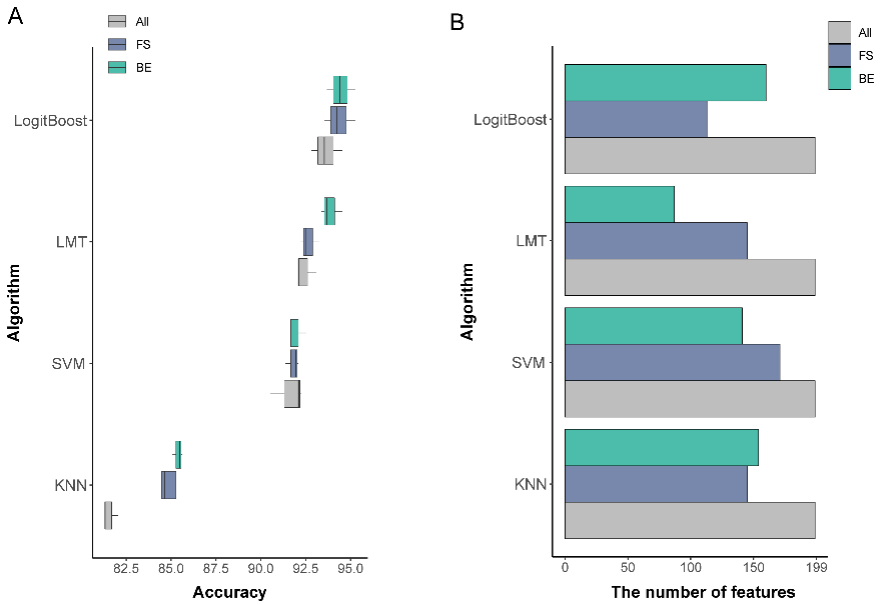


Figure 2.3. Classification performance by four classifiers at the genus level.

(a) Accuracies of four classifiers with three feature selection strategies (without feature selection, FS and BE). Evaluation of performance of each model involving different feature selection strategies was conducted three times. **(b)** The number of features by four classifiers with three feature selection strategies.

2.4.4 Accuracy, false positive and false negative error rate per six diseases

We examined the classification performance by calculating the accuracy of false positive rate (FPR) and false negative rate (FNR), which is a calculation method used to classify into two classes(Roberfroid, Gibson et al. 2010). Additionally, we investigated the performance of classification per diseases by obtaining feature set from BE with the highest performance. In LogitBoost algorithm, which had the highest performance among classifiers, average accuracies by disease was 98.1%, which is higher than overall accuracy of BE (93.6%) (Table 2.2). This increase of accuracy was caused by a higher number of true negatives because we applied calculation for evaluating a binomial classification for each disease. For the same reason, the mean of FPR (1.26%) was lower than that of FNR (13.86%). Since FPR divides true positive by sum of a true negative and true positive which makes it inversely proportional to true negative, in our case, as the number of true negative jumps to a greater number, a lower value of FPR was observed. Out of six diseases, CRC showed the highest FPR (3.7%) of all the diseases, which implies the classification of 3.7% of patients with non-CRC diseases as CRC. The lowest accuracy in CRC (96.84%) among six diseases was caused by a highest FPR. As FNR of the diseases showed high variance between diseases, CRC, HIV1, and Stroke (2.28, 0.36, 3.78%) were less than 5% of FNR, whereas JIA, ME/CFS, and MS (16.09, 28.47, 32.18 %) were more than 10 % of FNR. Diseases with high FNR including JIA, ME/CFS, and MS showed higher occurrences of misclassification into other diseases. In contingency tables, we observed that

diseases with a high FNR are highly likely to be classified as CRC which had the highest FNR of all diseases.

The diseases with high FPR and FNR in other algorithms were the same as that in LogitBoost algorithm. CRC had the highest FPR and the lowest accuracy among diseases in other classifiers. JIA, ME/CFS, and MS had higher FNR compared to other diseases in other classifiers. In KNN algorithm, CRC showed the highest FPR of 12.93%, while other classes showed FPR lower than 3%. Also, FNR of JIA, ME/CFS and MS (34.48, 64.58 and 77.01%) were higher than that of other classes with FNR below 8%. However, classes with higher FPR (or FNR) in KNN showed higher FPR(or FNR) compared to that in LogitBoost. FPR of CRC in KNN (12.93%) is three times higher than that in LogitBoost (3.7%). FNR of JIA, ME/CFS and MS (58.69%; mean of three classes) in KNN is twice as much as that in LogitBoost (25.58%; mean of three classes).

Table 2.2 Evaluation of performance per class in feature subset of BE in four algorithms.

The model was validated by 10-fold cross-validation and repeated three times. Values represent the mean of accuracy ± variance.

Accuracy							
	CRC	HIV1	JIA	ME/CFS	MS	Stroke	Average
LogitBoost	**96.84±0.43**	99.71±0.14	98.52±0.22	96.93±0.46	98.28±0.29	98.32±0.46	98.1±0.33
LMT	**95.93±0.3**	98.66±0.22	98.95±0.22	96.26±0.57	98.18±0.44	98.8±0.22	97.8±0.33
SVM	**95.59±0.5**	98.85±0.25	98.28±0.38	96.46±0.08	98.08±0.22	98.75±0.22	97.67±0.28
KNN	**90.28±0.3**	97.27±0.43	97.27±0	94.73±0.36	96.41±0.14	96.55±0.5	95.42±0.29
FPR							
	CRC	HIV1	JIA	ME/CFS	MS	Stroke	Average
LogitBoost	**3.7±0.83**	0.26±0.11	0.85±0.09	1.18±0.24	0.4±0.09	1.14±0.28	1.26±0.27
LMT	**3.93±0.4**	0.85±0.3	0.6±0	1.7±0.41	0.7±0.43	0.9±0.18	1.45±0.29
SVM	**4.77±0.48**	0.59±0.2	0.8±0.09	1.59±0.09	0.9±0.15	0.6±0.1	1.54±0.19
KNN	**12.93±0.23**	1.83±0.49	1.35±0.15	0.87±0.32	0.4±0.17	2.34±0.18	3.29±0.26
FNR							
	CRC	HIV1	**JIA**	**ME/CFS**	**MS**	Stroke	Average
LogitBoost	2.28±0.38	0.36±0.31	**16.09±3.98**	**28.47±9.62**	**32.18±7.18**	3.78±1.48	13.86±3.82
LMT	4.31±0.22	2.69±0	**11.49±5.27**	**31.25±3.61**	**27.59±3.45**	2.36±1.64	13.28±2.37
SVM	3.8±0.66	2.69±1.08	**22.99±7.18**	**29.86±1.2**	**25.29±1.99**	3.78±0.82	14.74±2.16
KNN	4.44±0.44	5.2±0.31	**34.48±3.45**	**64.58±2.08**	**77.01±5.27**	7.8±2.56	32.25±2.35

2.4.5 Identification of the disease-related microbial features

Through feature selections, we detected feature subsets that distinguish six diseases with the highest performance per classifier. Selected features can be used for microbial marker as they may be a shred of evidence of a close relatedness with the six diseases(Walker, Ince et al. 2011). Thus, we predicted that our selected features can also be applied as biomarkers for the six diseases. Among the potential biomarkers, we examined commonly selected genus in eight selected feature subsets at the genus level from the multiplication of four classifiers and two feature selection methods. The number of common selected features in FS and BE were 94, 66, 120 and 116 in LogitBoost, LMT, SVM, and KNN algorithm, respectively (Supplementary Figure 2.1). Among them, 17 genera were commonly identified in all four classifiers (Table 2.3). To elucidate further on the importance of these genera in classification, we looked closely into the rank of individual genus. The rank of the genus to be added or dropped during the feature selection procedure could be of interest as the features with greater performance tends to be added earlier or dropped later during feature selection. Therefore, we considered the rank of genus in the selection. Among 17 genera, only PSBM3 was selected in order of no more than five, which is less than 5% of 199 genera (10 number of genera). PSBM3 belongs to a bacterial family called Erysipelotrichaceae, which is associated with immune system(Oksanen and Blanchet). Erysipelotrichaceae was coated by IgA and their abundance had a positive correlation with tumor necrosis factor alpha levels(Dinh, Volpe et al. 2014, Callahan, McMurdie et al. 2016). Specifically, PSBM3 is associated with invariant natural killer T, which had a crucial role in pathogenesis of inflammatory diseases(Zhang, Kobert et al. 2014)

Table 2.3 Robust genera subset from two feature selection methods in four classifiers.

We present 17 genera selected in combination of four classifiers and two feature selection method. Column represent "Classifier / feature selection method". The figures in the table show the order of genera in selection steps. The lower number (figure) indicates the more importance for genera in terms of performance.

	Logit Boost/FS	LogitBoost/BE	LMT/FS	LMT/BE	SVM/FS	SVM/BE	KNN/FS	KNN/BE	Mean of order
PSBM3	3	2	5	3	3	2	3	3	3
Candidatus Azobacteroides	6	10	7	8	10	122	5	60	28.5
Cetobacterium	10	19	6	25	19	31	17	154	35.125
Ralstonia	46	17	93	14	27	16	45	24	35.25
Proteus	32	3	126	15	6	27	9	78	37
Flavobacterium	33	7	98	51	44	17	49	7	38.25
Moryella	8	105	1	77	7	1	103	65	45.875
Citrobacter	11	89	20	5	88	7	135	13	46
Anaerofustis	23	6	35	73	66	26	129	36	49.25
Dickeya	18	26	27	10	171	11	28	111	50.25
Owenweeksia	52	16	95	6	8	131	68	58	54.25
Salmonella	22	69	99	61	49	59	125	77	70.125
Pediococcus	99	93	46	82	67	45	145	19	74.5
Variovorax	80	127	54	79	133	79	58	57	83.375
Leuconostoc	83	112	96	63	63	91	94	88	86.25
Marvinbryantia	106	156	118	43	80	113	78	89	97.875
Novosphingobium	51	151	121	48	90	82	116	151	101.25

2.5 Discussion

We compared the performance of classification for six diseases in terms of three factors: 1) taxonomy level, 2) classifier and 3) feature selection method. Among the three factors, altering taxonomy levels influenced the classification performance the most. Moreover, we found that the performance improved as we used lower taxonomy level as features, which is consistent with a previous finding(Ozdal, Sela et al. 2016). Microorganisms at lower taxonomy levels have been used to investigate their impact on the host because they help to estimate the function more specifically(Tan, Wang et al. 2018). This suggests the necessity of using the technology of assigning microorganisms with high resolution in the classification of various diseases. In addition to the taxonomy level, we also evaluated the classification performance of four classifiers. Among the four classifiers, LogitBoost showed the highest performance. LogitBoost algorithm is a boosting model which process interactions effectively and robust to outliers, missing data, and many correlated as well as less important variables(Ravussin, Koren et al. 2012, Goodrich, Waters et al. 2014, Lee, Kim et al. 2014, Hwang, Lee et al. 2015) . This might have a positive influence on enhancing the performance of the classification of multiple diseases. On the other hand, KNN showed the lowest performance. KNN algorithm is reasonably well solved for a smaller number of features(Conlon and Bird 2014). The performance of KNN algorithm was especially lower at the genus level compared to the other classifiers.

We constructed feature subsets using FS and BE. FS and BE achieve improved accuracy because they find the optimal feature sets by interacting with classifiers(Roh, Kwon et al. 2016). On the other hand, FS and BE require

expensive computation times with a large number of features. This might rarely cause their application in the gut microbiome data. In this study, we showed that the selected microorganisms with FS and BE could boost the performance, especially, the feature subsets selected by BE had higher performance than that by FS. Since BE starts with the full set of features, it is easier to capture the interactive features, such that this advantage of BE can take into account the complex network of microbe-microbe interactions. Microbes interact each other by forming microbial guilds where they provide the substrate to each other, and even some anaerobic bacteria in the gut were demonstrated to perform metabolic cross-feeding(Singh, Chang et al. 2017). Therefore, a group of microorganisms is more related to human health than individual ones, which is why a higher performance of BE was observed.

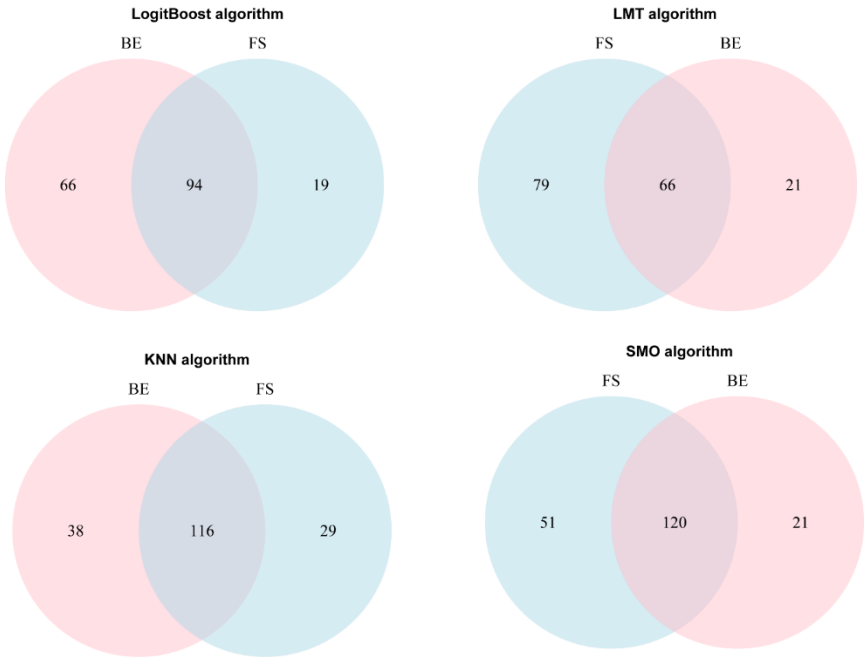
While performing the feature selection, we proposed the feature subsets that are potentially related to six different diseases. However, the feature subsets selected in this study may not contain all the microorganisms associated with the six diseases due to the data preprocessing. We preprocessed the data with various measures such as employing strict criteria when collecting the data from various studies and performing TMM and quantile normalization to minimize the variations between the studies. In addition, the samples were composed of a variety of nationalities which influence dietary habits, thereby affecting the composition of the gut microbiome. Some features, which might be affected by variation among samples, were deleted to reduce heterogeneity across different studies, which might cause by the effect of nationality. Thus, a few features significantly related to the six diseases may have been removed from this process. Despite the limitation of data preprocessing from different studies, we detected microorganisms associated with the six diseases.

Association with gut microbiome and health suggested the potential roles of gut microorganisms in precision medicine approach(Kashyap, Chia et al. 2017). Disease-related microorganisms can be used as microbial markers to detect diseases using well-known methods including metagenomics, phylogenetic microarrays, DNA fingerprinting techniques, and qPCR(Zhang, Kobert et al. 2014). Most of the previous disease studies on metagenome data focused on identification of biomarkers by comparing two groups of samples (case-control study)(Zeller, Tap et al. 2014). However, focusing on one disease may not be able to detect biomarker bacteria that is specific to that disease. This is because the same microorganisms can be differentially abundant in several diseases since the immune system of the host is influenced by the certain gut microbiome community that can be vulnerable to various diseases(Chang, Wang et al. 2011, Kinross, Darzi et al. 2011). On the other hand, the selected features in this study is expected to have disease specific profiling of microbial communities, which can be used for biomarkers to distinguish various diseases simultaneously. For example, PSBM3 (belongs to Family Erysipelotrichaceae) was an important feature in eight feature subsets. In the previous study, family Erysipelotrichaceae was studied to be associated with host diseases such as inflammatory bowel disease and HIV, as well as with the immune system(Oksanen and Blanchet , Dinh, Volpe et al. 2014, Callahan, McMurdie et al. 2016). This implies that the abundance of family Erysipelotrichaceae (or genus PSBM3) is an important clue to detecting multiple diseases.

As a result of the classification per diseases investigation, we found that JIA, ME/CFS and MS are classified into CRC. According to previous studies, CRC is related to fatigue symptom, which is a similar symptom with ME/CFS(Kim, Ju et al. 2017). The fatigue by CRC can be affected by

sarcopenia, characterized by muscle loss, which demonstrates the relationship between ME/CFS and CRC(Sainani, Desai et al. 1979). Moreover, there is a possible relationship between cancer risk and MS, which can cause diagnostic neglect(Ferrario, Taverniti et al. 2014). Though, the association between CRC and JIA has not been identified.

In summary, we presented the classification of six diseases using a machine learning algorithm and gut microbiome data. By evaluating performance in various perspectives, we showed the effect of bacterial abundance of different taxonomy levels and various classifier on performance of classification. Furthermore, we suggested the optimal genus subsets that can be potentially used as microbial markers to distinguish multiple diseases through feature selection, which confers the potential use for multi-diseases classification in the diagnosis of diseases.



Supplementary Figure 2. 1. The number of intersect of selected features from FS and BE

Supplementary Table 2. 1. Parameters for KNN algorithm.

The highest accuracy was shown in bold for each taxonomy level.

Parameter (K)	Accuracy				
	Phylum	Class	Order	Family	Genus
3	**53.88**	**66.95**	**71.12**	70.40	81.75
5	52.73	65.80	70.98	68.82	81.75
7	53.30	64.22	69.54	69.11	**82.18**
9	53.45	63.07	69.54	70.11	81.47
11	53.16	62.21	69.25	**71.12**	82.18
13	53.30	61.93	69.11	70.55	81.61
15	53.88	60.20	67.10	70.40	80.75

Supplementary Table 2. 2. Parameters of LogitBoost algorithm.

The highest accuracy was shown in bold for each taxonomy level.

Parameter (I)	Accuracy				
	Phylum	Class	Order	Family	Genus
1	52.16	60.34	65.52	71.70	77.73
2	53.30	64.37	69.40	71.70	81.18
3	55.03	67.53	71.12	77.59	84.63
4	**57.18**	67.53	71.55	78.45	85.92
5	55.75	68.53	75.14	79.89	88.07
6	54.60	70.11	74.86	80.75	91.09
7	55.17	70.40	76.29	81.47	90.66
8	55.03	69.97	76.15	82.04	91.24
9	54.60	70.26	77.59	81.61	91.95
10	55.60	70.69	76.01	82.61	91.95
11	55.03	70.98	77.01	82.90	92.82
12	54.60	70.98	77.87	83.05	92.67
13	55.32	70.69	76.72	82.18	92.67
14	54.60	70.83	77.73	83.19	93.39
15	54.45	70.98	77.59	82.47	93.82
16	54.74	70.55	78.02	83.48	92.82
17	53.88	70.69	77.44	84.05	94.54
18	54.89	70.83	78.02	83.91	92.67
19	54.89	70.40	76.87	83.91	93.97
20	54.89	71.12	78.45	85.06	93.97
21	54.60	70.69	78.45	84.77	94.54
22	55.17	71.12	77.44	84.77	94.68
23	54.60	**71.55**	77.87	84.63	94.54
24	55.03	70.40	**79.31**	84.20	94.40
25	54.89	70.83	79.02	84.48	93.39
26	54.60	70.69	79.31	84.34	93.53
27	55.32	71.12	77.16	83.62	94.25
28	54.60	70.83	77.87	84.77	93.97
29	54.45	70.98	77.16	84.20	94.68
30	55.75	69.83	77.87	**85.49**	**95.40**
31	54.89	71.12	77.73	84.48	94.40
32	54.89	70.26	78.02	84.48	95.11
33	54.45	70.55	78.16	84.63	93.97
34	55.46	70.83	78.45	83.33	93.82
35	54.74	70.98	77.87	84.05	94.68
36	55.03	70.55	78.45	84.63	94.11
37	55.32	70.83	78.59	82.90	94.97
38	54.45	70.40	79.02	84.77	93.97
39	54.89	70.69	79.02	84.63	94.11
40	54.89	70.98	77.30	83.48	94.83

Supplementary Table 2. 3. Parameter of RBF Kernel in SVM algorithm.

The columns represent parameter G, while the rows represent parameter C. The highest accuracy was shown in bold for each taxonomy level.

	Phylum					
	1e.04	0.001	0.01	0.1	1	10
0.1	37.79	37.79	37.79	52.59	54.31	55.03
1	37.79	37.79	52.44	54.89	54.17	**55.89**
10	37.79	50.86	54.89	54.60	54.74	54.45
100	52.44	54.89	55.60	55.32	55.03	54.74
150	53.45	54.89	55.17	54.45	54.17	55.03
200	53.59	54.89	55.03	54.74	55.03	54.89
300	53.74	54.89	55.32	54.89	55.32	54.89
400	54.31	54.89	54.31	54.45	54.17	54.45
1000	54.89	55.03	55.32	54.60	55.17	54.31
	Class					
	1e.04	0.001	0.01	0.1	1	10
0.1	37.79	37.79	37.79	54.60	64.66	48.85
1	37.79	37.79	56.18	63.79	68.68	68.97
10	37.79	56.18	61.49	69.11	69.25	68.68
100	56.32	61.35	68.97	70.40	69.40	68.68
150	59.63	68.25	69.11	70.11	70.26	68.53
200	60.34	68.68	69.97	70.40	69.97	68.82
300	60.34	68.25	70.11	69.97	69.83	68.97
400	60.78	68.82	70.26	69.68	69.40	68.82
1000	62.50	68.68	**70.55**	69.54	70.11	68.39
	Order					
	1e.04	0.00	0.01	0.10	1.00	10.00
0.1	37.79	37.79	37.79	51.58	69.40	41.38
1	37.79	37.79	60.20	74.57	78.02	59.91
10	37.79	61.64	75.43	77.87	75.86	61.93
100	59.63	74.43	**79.31**	76.87	74.86	61.78
150	68.97	76.29	77.87	78.59	76.15	61.49
200	69.40	76.29	78.02	78.45	76.01	63.22
300	71.70	77.73	77.16	78.02	76.58	61.49
400	72.41	78.16	77.73	78.02	75.14	62.36
1000	75.29	78.30	78.59	76.87	75.86	61.78
	Family					
	1e.04	0.00	0.01	0.10	1.00	10.00
0.1	37.79	37.79	37.79	49.43	51.58	37.79
1	37.79	37.79	59.91	77.73	78.30	38.22
10	37.79	62.07	78.88	81.90	78.30	39.22
100	61.49	78.59	82.61	79.45	79.17	38.94
150	68.68	79.17	82.04	80.17	78.30	39.08
200	70.98	79.17	81.90	77.59	79.89	38.79
300	74.14	80.60	**82.76**	79.74	79.17	38.79
400	75.57	81.03	82.61	79.45	79.02	39.22
1000	78.45	82.61	81.75	80.17	79.45	38.94
	Genus					

	1e.04	0.00	0.01	0.10	1.00	10.00
0.1	37.79	37.79	37.79	62.21	39.08	37.79
1	37.79	37.79	69.11	86.06	79.31	37.79
10	37.79	69.40	1.00	91.52	81.18	37.79
100	69.54	88.22	92.53	91.09	81.47	37.79
150	77.30	90.23	91.81	92.24	81.75	37.79
200	79.02	91.52	91.95	92.53	82.04	37.79
300	80.75	91.09	92.10	91.67	81.18	37.79
400	82.76	91.81	93.10	91.95	80.46	37.79
1000	87.21	92.53	**93.25**	92.39	81.32	37.79

Supplementary Table 2. 4. Selected parameter in disease classification.

	KNN	LogitBoost	SVM	
	K	I	C	G
Phylum	3	4	1	10
Class	3	23	1000	0.01
Order	3	24	100	0.01
Family	11	30	300	0.01
Genus	7	30	1000	0.01

This chapter will be published elsewhere
as a partial fulfillment of Sohyun Bang's M.Sc program.

Chapter 3. Effects of *Allium hookeri* leaf on gut microbiome related to growth performance of young broiler chickens

3.1 Abstract

Healthy food promotes beneficial bacteria in the gut microbiome. A few prebiotics acting as food supplements increase fermentation by beneficial bacteria, which enhance the host immune system and health. *Allium hookeri* belongs to Allium, which is considered healthy food with antioxidant and anti-inflammatory activities. *A. hookeri* as a feed supplement for broiler chickens is known to improve growth performance. Although the underlying mechanism has yet to be elucidated, it is suspected that it alters the gut microbiome. In the current study, 16S rRNA sequencing has been carried out using the samples from the cecum of broiler chickens exposed to a diet comprising different tissue types (leaf and root) and amounts (0.3% and 0.5%) of *A. hookeri* to investigate their impact on gut microbiome. The composition of microbiome in the group supplemented with *A. hookeri* leaf 0.5% was distinguished from that in the control group. Differences in microbial communities included decreases in the abundance of bacteria in some genera. The modulated gut microbiome by leaf 0.5% supplement was correlated with growth traits including body weight, bone strength, and infectious bursal disease antibody. The results demonstrate that the benefit of *A. hookeri* for broiler chicken is related to the altered gut microbiome.

3.2 Introduction

Diet plays an important role in modulating gut microbiome by providing food substrates for gut microorganisms (Conlon and Bird 2014, Kim, Lillehoj et al. 2015). Some dietary components not digestible by the host enzymes can be digested by gut bacteria (Gerritsen, Smidt et al. 2011). For example, prebiotics such as inulin, polyphenol, and galacto-oligosaccharide are non-digestible food ingredients that can promote the growth of beneficial bacteria (Ferrario, Statello et al. 2017). Such prebiotics increase the fermentation products produced by beneficial bacteria, which enhances host immune response (Roberfroid, Gibson et al. 2010, Rooks and Garrett 2016). Elucidation of the interactions between diet and microbiome has raised the interest in functional foods with beneficial effects on gut microbiome and host health (Singh, Chang et al. 2017).

Among the various functional foods, *Allium hookeri* is widely used as a healthy food that treat high blood glucose or lipid levels in patients with diabetes mellitus in Korea (Kim 2015, Lee 2015). *A hookeri* belongs to Allium, a widely known genus that contains onion (Allium cepa) and garlic (Allium sativum L.). Plants of genus Allium have been used as medicinal foods to reduce the risk of several types of cancers by preventing mutagenesis (Sengupta, Ghosh et al. 2004). The beneficial effects of Allium are attributed to the abundance of organosulfur compounds, polyphenols, and allicin (Bianchini and Vainio 2001, Nencini, Cavallo et al. 2007). *A. hookeri* contains six-fold higher levels of organosulfur than garlic, and higher cellulose and total phenol contents than onion (Kim, Lee et al. 2012). As these components exhibit antioxidant activities, the use of *A. hookeri* as a medical food is promising (Lee, Kim et al.

2014, Hwang, Lee et al. 2015). *A. hookeri* has shown immunomodulant effects in lymphocytes, macrophages, and tumor cells in in-vitro chicken cell experiments (Lee, Lee et al. 2017). In vivo experiments have also suggested that *A. hookeri* inhibits the inflammatory response in the pancreas of diabetic rats and LPS-induced young broiler chickens (Roh, Kwon et al. 2016, Lee, Lee et al. 2017).

The beneficial effects of *A. hookeri* on health suggest its use in commercial animal farming including pigs and chickens (Viveros, Chamorro et al. 2011, Song, Pyun et al. 2014, Kim, Ju et al. 2017). *A. hookeri*, when used as a feed supplement, enhanced the oxidative stability of pork and improved the growth performance of broiler chickens (Eun Byeol Lee). Although the mechanism of action is unclear it is suspected that *A. hookeri* alters the gut microbiome. *A. hookeri* components such as organosulfur compounds, polyphenols, and allicin are known to affect the gut microbiome by increasing or decreasing the components associated bacteria (Fujisawa, Watanabe et al. 2009, Viveros, Chamorro et al. 2011, Carbonero, Benefiel et al. 2012). Moreover, a previous study has reported that diets including onions belonging to Allium genus modulate gut microbiota and increase body weights of broiler chickens (LEE 2013, ur Rahman, Khan et al. 2017). Thus, it has been hypothesized that *A. hookeri* alters the gut microbiome and that such changes might lead to beneficial growth effects in commercial animals27. However, the collective effect of *A. hookeri* on gut microbes needs to be further elucidated.

Thus, the objective of this study was to determine the effect of *A. hookeri* as a feed supplement on chicken gut microbiome based on 16S rRNA sequencing. We sequenced a total of 24 caecal samples derived from six groups of chickens (four samples each). The groups include chickens fed with 0.3%

leaves, 0.5% leaves, 0.3% roots and 0.5% roots. Effects of *A. hookeri* as a feed additive were evaluated by comparing the *A. hookeri* diet groups with the control group (Control) or commercial supplement group (CS). The microbial diversity and abundance of *A. hookeri* groups were compared with those of Control and CS.

3.3 Materials and methods

3.3.1 Sample collection

A total of 1,200 male broiler chickens (Arbor Acres broilers) were grown for 35 days. They were divided into six groups (n=200 chickens/group): Control, CS, Leaf 0.3, Leaf 0.5, Root 0.3, and Root 0.5. All groups were freely fed with a basal diet (crude protein (CP) 22%, metabolic energy(ME) 3,100 kcal/kg for 0-3 weeks and CP 20%, ME 3,150 kcal/kg for 3-5 weeks). Commercial Xtract (ML Co, Seoul, Korea) was used at 0.05% of the diet in the CS group. *A. hookeri* Leaf or root powder (0.3% or 0.5%) was added to Leaf 0.3, Leaf 0.5, Root 0.3, and Root 0.5 groups. The different powders of *A. hookeri* were prepared after freeze drying and grinding. Four chickens in each of the six groups with their similar to the average group weight were selected. Thus, a total of 24 samples (four samples per each group from six groups) were used for sequencing.

3.3.2 DNA extraction and Illumina Sequencing

Cecal samples were used for DNA extraction using AccuPrep Stool DNA Extraction Kit following the manufacturer's instructions. The V3 and V4 region of the 16S rRNA genes was PCR amplified from the microbial genomic DNA. The DNA quality was determined by PicoGreen and Nonodrop. The input gDNA (10 ng) was PCR amplified using the barcoded fusion primers 341F/805R (341F: 5' CCTACGGGNGGCWGCAG 3, 805R: 5' GACTACHVGGGTATCTAATCC 3'). The final purified product was quantified using qPCR according to the qPCR Quantification Protocol Guide

(KAPA Library Quantification kits for Illumina Sequencing platforms) and qualified using the LabChip GX HT DNA High Sensitivity Kit (PerkinElmer, Massachusetts, USA). The 300 paired-end sequencing reaction was performed on MiSeq™ platform (Illumina, San Diego, USA). The sequencing data were deposited into the Sequence Read Archive (SRA) of NCBI (http://www.ncbi.nlm.nih.gov/sra) and can be accessed via accession number SRP151247.

3.3.3 Taxonomic analysis

Demultiplexed paired-end reads were merged with PEAR (Zhang, Kobert et al. 2014). Pre-processed reads were analyzed using QIIME2 version 2017.12. We used DADA2 software package (Callahan, McMurdie et al. 2016) implemented in QIIME2 to model and correct Illumina sequenced FASTAQ files by removing chimeras using "consensus" method. QIIME2 q2-feature-classifier plugin was trained on Silva database (Release 128) for 99% OTU full-length sequences.

Alpha and beta-diversity analyses were performed with q2-diversity plugin in QIIME2 at a sampling depth of 1000. Weighted Unifrac distance matrix was used for Permutation multivariate analysis of variance (PERMANOVA) and PCoA plot. PERMANOVA was performed with 999 permutations to weighted UniFrac distance matrix using Adonis in R package 'vegan' (Oksanen and Blanchet).

3.3.4 Identification of differentially abundant microbiomes (DAM)

Trimmed Mean of M values (TMM) was obtained to adjust for different library sizes using edgeR (Robinson, McCarthy et al. 2010). Statistical tests were performed under generalized linear model (GLM) considering OTU's count as negative binomial distribution. To compare the goodness-of-fit of two models, the log-likelihood ratio statistic was calculated. In the statistical test, the false discovery rate (FDR) was used to adjust for multiple testing error with a significance level of 5% (Benjamini and Hochberg 1995).

3.3.5 Correlation between microbiota and growth traits

The Spearman correlation was used between TMM values of each sample and growth traits (body weight, bone strength and IBD antibody). The abundance (TMM value) of significantly correlated genera was visualized in heatmap using pheatmap R package.

3.3.6 PICRUSt analysis and statistical comparison of functions between groups

Phylogenetic investigation of communities by reconstruction of unobserved states (PICRUSt) was used to predict functional profile of microbiota (Langille, Zaneveld et al. 2013). Since PICRUSt uses a closed reference OTU picking based on Greengenes database (version 13.5.), the features assigned to Greengenes databases were used. The abundance of functions in Control was compared to that in Leaf 0.5 using Wilcoxon rank sum test.

3.4 Result

3.4.1 Effects of *A. hookeri* as a feed supplement on gut microbiome diversity

To elucidate the differences in the microbiota exposed to different amounts of *A. hookeri* roots or leaves, a principal coordinate analysis (PCoA) based on weighted UniFrac metrics was performed (Figure 3.1). Samples within the leaf group were clustered with shorter distance compared with those in the other groups. Permutational Multivariate Analysis of Variance (PERMANOVA) analysis was also performed to determine significant differences between groups. The variability between the six groups was observed (P-value < 0.05; Table 3.1). We also examined the effect of factors (leaf and root of *A. hookeri*) in distinguishing gut microbial communities by grouping the samples exposed to *A. hookeri* according to the tissue or amount. When we divided the samples exposed to *A. hookeri* to leaves and roots irrespective of the amounts of *A. hookeri*, significant differences were detected between Control, CS, Leaf and Root. On the other hand, we did not find significant differences between Control, CS, 0.3% and 0.5%, when the samples were divided according to the amount of *A. hookeri*.

PERMANOVA pair-wise tests were conducted for the groups exhibiting signifciant variation: (1) groups of 'Tissue of *A. hookeri'*, and (2) 'Tissue and amount of *A. hookeri'*. In the pair-wise test for (1), 'Control, Root' showed no significiant difference, although the difference between control groups and leaf was significant, with a low p-value (Table 3.2). In the pair-wise test for (2), pairs of groups exposed to the same tissues but different amounts (Leaf 0.3, Leaf 0.5 pair and Root 0.3, Root 0.5 pair) showed no significant difference

(Supplementary Table 3.1), which suggested that the effect of 0.3% and 0.5% *A. hookeri* showed little difference as in-feed supplements. Combinations with Leaf 0.5 were significant with lower p-values than other pair-wise combinations. This indicates that the abundance and variety of microorganisms in leaf 0.5 were different from those in the control group.

Microbial diversity within a local community was evaluated based on richness and diversity using the observed operational taxonomic unit (OTU) and Shannon index, respectively (Supplementary Figure 3.1). Richness showed similar distribution across groups with 'combinations of tissue and amount' of *A. hookeri* ($P = 0.083$) except for Leaf 0.3 compared with CS ($P = 0.02$). The diversity also showed a similar distribution of Shannon index across combinations except for Leaf 0.3 and Root 0.3 compared with CS ($P = 0.02$). Compared with CS, richness and diversity were not affected by diet supplemented with *A. hookeri* ($P = 0.5$). In summary, while the differences between Leaf and other groups showed better diversity, the differences within a local community were not apparent when compared with the control.

Table 3.1 Result of PERMANOVA

Test group	r^2	F-Ratio	P-value
Tissue and amount of *A. hookeri*	0.27	1.34	0.016
Tissue of *A. hookeri*	0.17	1.44	0.011
Amount of *A. hookeri*	0.13	1.07	0.31

Table 3.2. Results of PERMANOVA of Pair-wise test for tissue groups

Pair	P-value	Pair	P-value
Control, CS	0.537	CS, Leaf	0.018
Control, Leaf	0.022	CS, Root	0.438
Control, Root	0.056	Leaf, Root	0.005

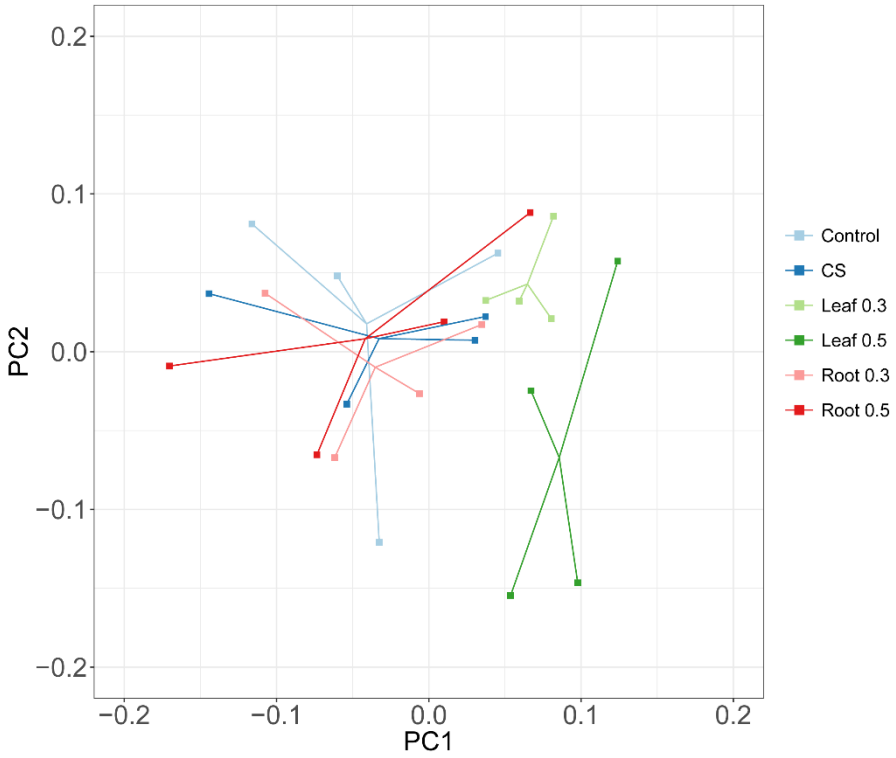


Figure 3.1. A principal coordinate analysis (PCoA) plot showing dissimilarities among different diet groups.

PCoA from distance of weighted UniFrac. Each dot represents one sample. “Leaf”, “Root”, “0.3”, “0.5” denote leaf, root of *A. hookeri*, 0.3%, and 0.5% in each diet, respectively.

3.4.2 A. *hookeri*-associated microbiota changes

At the phylum level, Firmicutes and Bacteroidetes were major microbes in the cecum of chicken (Figure 3.2a), accounting for more than 80% and 10% of microbes, respectively. Most phyla including Firmicutes and Bacteoidetes showed no difference between supplements and Control or CS except for Cyanobacteria identified as Gastranaerophilales at the class level, which was decreased in Leaf 0.5 compared with that of the Control or CS. However, we obtained only limited information related to the differences at the phylum level. Therefore, the differences in microbial abundance at the genus level were investigated.

Approximately 70% of the features have been identified at the genus level. These features were used to determine the differential abundance of microorganisms by comparing each *A. hookeri* supplement group with the Control or CS (FDR < 0.05). For the root supplement, no differential abundance of taxa was observed. On the other hand, a few genera were significantly associated with Leaf compared with the Control or CS. Leaf 0.5 carried had seven differential genera compared to the control and five compared to the CS, whereas leaf 0.3 had only two differential genera compared to the CS. Based on this finding, we concluded that only the Leaf supplement affected the abundance of microbial genera. The amounts of leaf supplement also affected the degree of altered microbial abundance.

The abundance of a few common genera varied in Control vs. Leaf 0.5, CS vs. Leaf 0.5, and CS vs Leaf 0.3. Parabacteriodes and [Eubacterium] nodatum group, which were differentiating genera in CS vs. Leaf 0.3 also varied in abundance in Control vs. Leaf 0.5. Most genera in Leaf 0.5 vs. CS were also identified in Leaf 0.5 vs. Control. [Eubacterium] nodatum group, Marvinbryantia, Oscillospira, and Gerlria were of common to both tests (Figure 3.2b). Significant and differentially abundant

genera detected only in Leaf 0.5 vs. Control included Parabacteroides, Gastranaerophilaes, and NB1-n while Christensenellacea R-7 group was differentially abundant only in Leaf 0.5 compared with CS. We observed a decreased abundance in most of the differentially abundant genera except for Parabacteroides (Figure 3.2c). Eubacterium nodatum group had almost zero abundance in both Leaf 0.3 and Leaf 0.5. The abundance of Marvinbryantia, Gelria and NB1-n decreased gradually according to the amount of Leaf supplemented compared with their mean abundance in the Control. A rapid reduction in the abundance of Oscillospira, Gastranaerophilales, and Christensenellaceae R-7 group was observed in Leaf 0.5, similar to the Control and Leaf 0.3 groups.

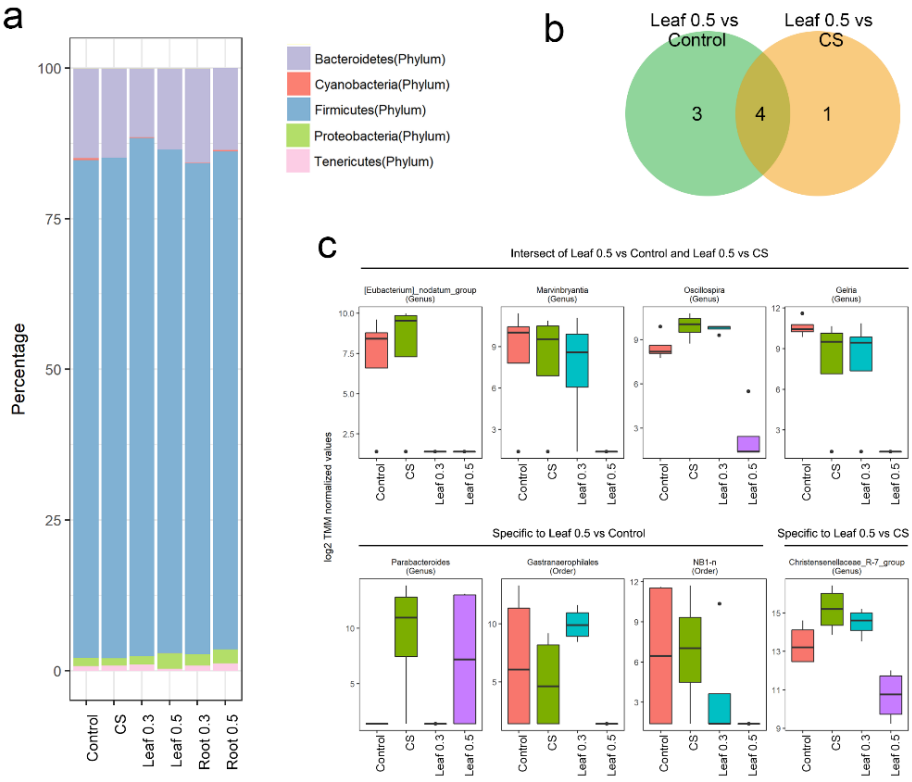


Figure 3.2 Abundance of microorganisms at phylum and genus levels.

(a) Phylum level composition. Bar plots represents the percentage (%) of average abundance for groups.

(b) Venn diagrams showing differentially abundant features at genus level.

(c) Relative abundance of differentially abundant taxa in Leaf.

3.4.3 Correlation between microbial abundance and growth traits

A. hookeri not only affects microbiome, but also the growth traits such as body weight, bone strength, and infectious bursal disease (IBD) antibody. These traits were linkedwith gut microbiome in several studies (Ravussin, Koren et al. 2012, D'Amelio and Sassi 2018). These associations were examined by correlation analysis of the abundance of genera and growth traits. In the association test for Spearman's rank correlation coefficient (Spearman's rho; P-value < 0.05), 7, 5, and 7 genera showed significant correlations with body weight, bone strength, and IBD antibody, respectively.

Among the seven genera correlating with body weight, Ruminococcaceae UCG-005 showed highly negative correlation with body weight (Spearman's rho = -0.56) (Figure 3.3a). Clostridium sensu stricto 1 had the highest correlation with body weight (Spearman's rho = 0.44). The relative abundance of correlated genera was indicated next to the correlation index. The genera showing negative correlation with body weight had lower abundance in Leaf 0.3 and Leaf 0.5 while the genera displaying positive correlation with body weight showed relatively higher abundance in Leaf 0.3 and Leaf 0.5. Thus, microorganisms correlating with body weight are related to Leaf diet.

Among genera that correlated with bone strength, most correlations were negative (Figure 3.3b). Most microorganisms that were negatively correlated with bone strength had lower abundance in Leaf 0.5 compared with those in other groups. NB1-n genera showing negative correlation with bone strength (Spearman's rho = -0.43) were significant lower in Leaf 0.5. On the other hand, Parasutterella only had a high positive correlation with bone strength (Spearman's rho = 0.74).

Seven microorganisms were correlated with antibody titers against IBD, including three that were negatively correlated and four that were positively correlated (Figure 3.3c). The Eubacterium brachy group had the highest correlation (Spearman's rho = 0.57), showing increased abundance in Leaf 0.3 and Leaf 0.5 but uniformly lower abundance in other groups. Parasutterella was positively correlated with bone strength and with IBD antibody production. Gelria was significantly decreased in Leaf 0.5 and showed negative correlation with IBD antibody (Spearman's rho = -0.44). Leaf 0.5 also showed the same pattern of microbial abundance correlated with IBD antibody, with a higher (or lower) abundance in Leaf 0.5 for positive (or negative) correlation with IBD antibody.

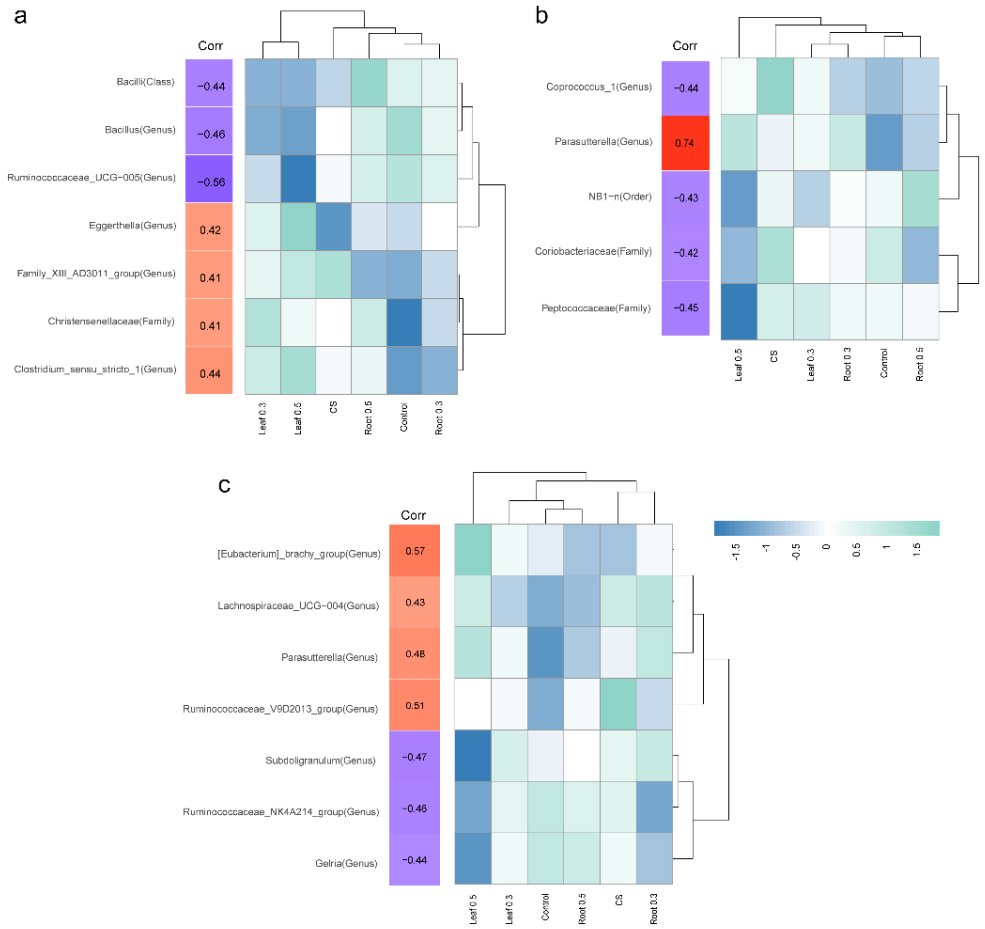


Figure 3.3. Correlation between microbiota and growth performance including body weight, bone strength, and IBD antibody.

(a) Microorganisms showing significant correlation with body weight. The column in “corr” shows Pearson's r value. The heatmap represents mean abundance of groups.

(b) Microorganisms showing significant correlation with bone strength.

(c) Microorganisms showing significant correlation with antibody titers against infectious bursal disease (IBD).

3.4.3 Prediction of gut microflora function

Microbial functions predicted by the phylogenetic investigation of communities via analysis of the reconstruction of unobserved states (PICRUSt) were compared between Control and Leaf 0.5 because the genera correlating with growth traits were associated with their abundance in Leaf 0.5 (Figure 3.4). There were 38 different functions; 30 have a lower abundance and eight have a higher abundance in leaf 0.5. Most higher functions in Leaf 0.5 were related to metabolism. Enriched functions in Leaf 0.5 such as C5-branched dibasic acid metabolism, fructose and mannose metabolism, and galactose metabolism were included in carbohydrate metabolism. Most significantly, lysine degradation was depleted in Leaf 0.5. Among the functions depleted in Leaf 0.5, the terms related to human disease were prevalent (Supplementary Figure 3.2). Especially, the depleted function in Leaf 0.5 included terms for cancers such as pathways in cancer, renal cell carcinoma, colorectal cancer, and small cell lung cancer.

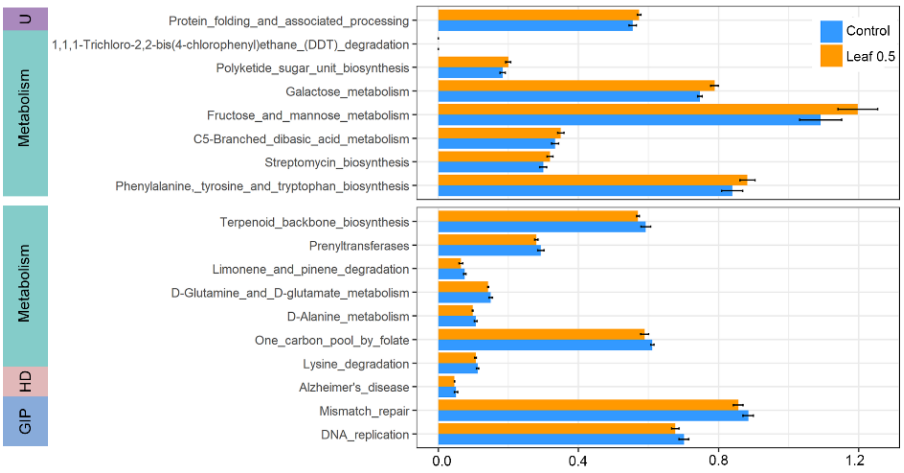


Figure 3.4. Predicted microbial functions showing significant difference between Control and Leaf 0.5.

Functions were predicted by PICRUSt. Functions with significant difference ($P < 0.05$, one sided Wilcoxon rank sum test) are shown. The error bar represents SD (standard deviation). The upper box shows functions enriched in Leaf 0.5 compared with control while the lower box shows functions enriched in control compared with Leaf 0.5 with top 10 p-value. U: Unclassified, HD: Human Diseases, GIP: Genetic Information Processing.

3.5 Discussion

This study showed the effect of *A. hookeri* on gut microbiome, especially in the cecum of broiler chicken. Towards this end, we fed different feed supplements including *A. hookeri* to six groups of 1,200 broiler chickens: Control, CS, Leaf 0.3, Leaf 0.5, Root 0.3 and Root 0.5, consisting of 200 chickens per group. Four chickens from each of six groups with their weights similar to the average weight of the group were selected. A total of 24 samples from six groups were used for sequencing. The small sample size is one of the study limitations; however, this is the first study to investigate the effect of *A. hookeri* on gut microbiome, despite several studies reporting its health benefits. Also, the impact of Leaf 0.5 on the altered gut microbiome in this study may corroborate our previous study, which highlighted improved growth traits in Leaf 0.5. Based on multiple analysis of gut microbiome and phenotype of broiler chickens, this study provides deeper insight into the relationship between the intake of *A. hookeri* Leaf, microbial communities and growth performance.

The remarkable effect of leaf on microbiome can be attributed to the abundance of components contained in leaves. Previous studies regarding *A. hookeri* have compared the concentrations of different components in leaf with those in the root. Compared with the root, the leaf contains a higher number of polyphenols, estimated by the amount of total phenolics and flavonoids (Hwang, Lee et al. 2015). Polyphenol is known to exhibit antioxidant activity that potentially prevents diseases related to oxidative stress, including cancer and cardiovascular diseases (Manach, Scalbert et al. 2004). A large proportion (~90%) of absorbable polyphenols are digested by gut microbiome rather than digestive enzymes in the small intestine (Edwards, Havlik et

al. 2017). During the degradation of polyphenols by microbiome, intermediates including aglycones that promote various aromatic acid metabolites are produced (Scalbert, Morand et al. 2002). The altered metabolites modulate the composition of microbial communities with prebiotic effects such as antimicrobial activities against pathogens (Lee, Jenner et al. 2006). In broiler chickens, polyphenol-rich grape increased Enterococcus but decreased Clostridium (Viveros, Chamorro et al. 2011). The effect of polyphenols in *A. hookeri* on microbiome might play an important role in modulating microbial communities, which explains the presence of unique microbial communities in polyphenol-rich leaf groups.

The altered abundance of bacteria following the supplementation of *A. hookeri* leaf occurs at the generic level. Among eight leaf-associated genera, the abundance of Parabacteroides was increased in Leaf 0.5. Parabacteroides are obligatory anaerobic bacteria that degrade saccharides, forming acetate and succinate as major end products (Sakamoto and Benno 2006). Except for Parabacteroides, the abundance of *A. hookeri*-associated genera was decreased. The reduction in abundance of most microbiome may be associated with the antibacterial effect of *A. hookeri* leaf. The decrease in bacterial concentrations may be attributed to allicin, which accounts for the spicy flavor of *A. hookeri*. Allicin is an organosulfur compound with characteristic antibacterial activity against a wide range of microorganisms including Staphylococcus and Pseudomonas in several studies (Cutler and Wilson 2004, Fujisawa, Watanabe et al. 2009). Its antibacterial activity is attributed to its chemical reaction with thiol groups of enzymes that influence the metabolism of cysteine proteinase activity related to bacterial virulence (Reiter, Levina et al. 2017). However, the mechanisms of antibacterial activity of *A. hookeri* are not fully understood.

Therefore, further evidence is needed to establish the relationship between the decreased microbial abundance and the antibacterial activity of *A. hookeri*.

The effects of *A. hookeri* on crucial phenotypes of boiler chickens were investigated, given their importance of growth trait enhancement without antibiotics for growth promotion (AGPs) and to decrease the risk of antibiotic resistance (Cervantes 2015). Other studies investigated feed supplements using natural foods including extracts from herb species, which enhanced poultry health by improving immunity and protecting chickens from avian diseases (Lillehoj, Kim et al. 2011, Kim, Lillehoj et al. 2013). Our previous study has shown that broiler chickens supplemented with *A. hookeri* show improved growth performance (Eun Byeol Lee). In the present study, a subset of samples from our previous study was used. Leaf 0.3 and Leaf 0.5 showed higher body weights (P-value: 0.001 and 0.052, respectively) compared with the control group (Supplementary Figure 3.3). To determine whether microbiome was related to greater body weight in leaf and to identify genera associated with body weights, the correlation between abundance of genera and body weight was investigated. Notably, genera showing positive (or negative) correlation with body weight were relatively abundant (or depleted) in Leaf 0.3 and Leaf 0.5, suggesting that alterations in microbiome induced by leaf may affect the body weight. Specific genera related to body weight are known to be related to diet or energy metabolism in other studies. For example, Bacillus that showed negative correlation with body weight was associated with feed efficiency in broiler chicks (Santoso, Tanaka et al. 1995). Clostridium sensu stricto 1 had positive correlation with body weight. Family Clostridiaceae is known to induce weight gain in rex rabbits (Zeng, Han et al. 2015). The association between increased body weight and microbiome is also supported by functional analysis. Results of functional analysis revealed that

Leaf 0.5 was more enriched in carbohydrate metabolism, which increased the body weight. In summary, the correlation between body weight and enrichment of carbohydrate metabolism provides evidence suggesting that altered microbial communities induced by *A. hookeri* may alter body weight.

The association between microbiome and other traits including bone strength and IBD antibody was also determined. Bone strength was examined because several bacteria improved calcium absorption (Ohlsson and Sjögren 2015). IBD antibody was investigated as an index of immune system related to microbiome (Carbonero, Benefiel et al. 2012). Among five genera associated with bone strength, none of them was investigated for correlation with bone strength. However, the genera associated with IBD antibody were consistent with or contrary to previous metagenome studies related to immune system or IBD. Lachnoclostridium and Ruminococcaceae groups that were positively and negatively correlated with IBD antibody, respectively, were also involved in the inoculation of very virulent IBD virus in broiler chickens (Li, Kubasová et al. 2018). This indicates that these genera are strongly linked to the regulation of immune system.

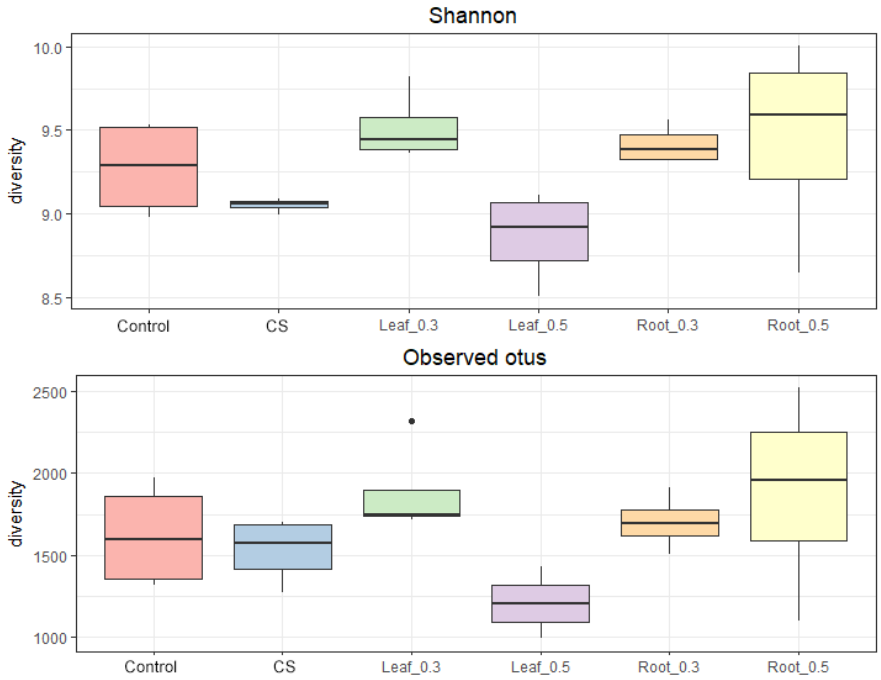
Parasutterella, which showed a positive correlation with bone strength was also positively correlated with IBD. The cause of common genera associated with bone strength and IBD might be explained by the relationship between immune and bone cells (D'Amelio and Sassi 2018). Their interaction is induced by bone marrow generating circulating blood cells containing T cell and B cell. T- and B-cells play important roles in cell-mediated immunity to regulate bone resorption and bone formation by producing large amounts of cytokines (Pacifici 2010), which explains the connection between bone and immune cells. A mouse study has revealed that gut microbiota regulate bone mass and immune status (Sjögren, Engdahl et al. 2012).

Despite the associations of specific genera with both traits, microorganisms known to control both bone mass and immune system were poorly detected. Thus, genera detected in this study may provide a clue for this interaction. Further studies are needed to determine the role of these genera in the interaction with immune system and bone strength.

In conclusion, our results showed the degree of alteration based on supplementation of *A. hookeri*, especially its leaf. Microbial communities were decreased in abundance in the leaf group indicating that the leaf has a distinct effect on microbial communities. We also detected specific genera related to body weight, bone strength, and IBD antibody known to be important for the productivity of broiler chicken. Therefore, the benefit of *A. hookeri* for broiler chicken is mediated via its microbiome, suggesting that *A. hookeri* has potential as a feed supplement for broiler chickens.

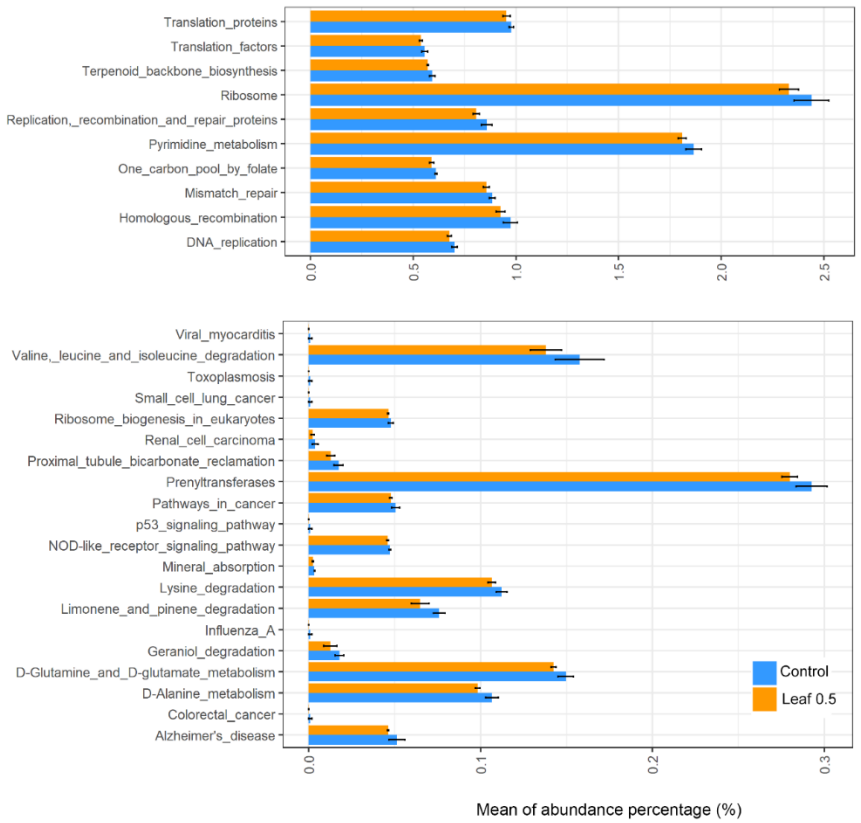
Supplementary Table 3. 1. Results of PERMANOVA of Pair-wise test for combinations groups

Pair	P-value	Pair	P-value
Leaf 0.3, Leaf 0.5	0.115	Leaf 0.5, Control	0.058
Leaf 0.3, Control	0.086	Leaf 0.5, CS	0.027
Leaf 0.3, CS	0.056	Leaf 0.5, Root 0.3	0.029
Leaf 0.3, Root 0.3	0.028	Leaf 0.5, Root 0.5	0.029
Leaf 0.3, Root 0.5	0.229	CS, Root 0.3	0.656
Control, CS	0.544	CS, Root 0.5	0.486
Control, Root 0.3	0.685	Root 0.3, Root 0.5	0.715
Control, Root 0.5	0.837		



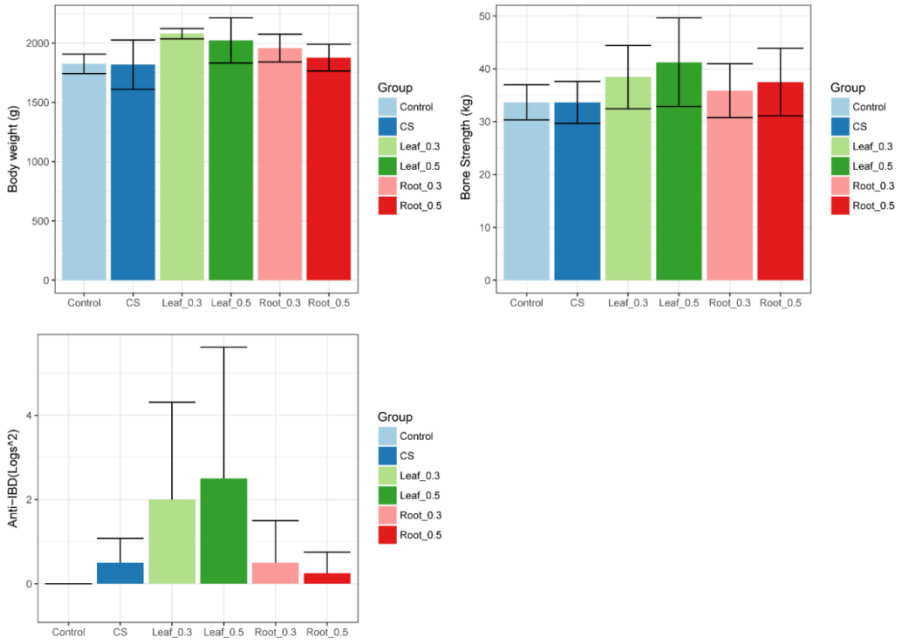
Supplementary Figure 3 1. Shannon index and observed OTUs in six groups.

In Shannon index, significant difference was observed in Leaf 0.3 and Root 0.3 compared to CS ($P = 0.02$). In observed OTU, Leaf 0.3 was significantly different from CS ($P = 0.02$).



Supplementary Figure 3 2. Predicted microbial functions deficient in Leaf 0.5.

Functions were predicted by PICRUSt. All functions enriched in Control compared to Leaf 0.5 were presented (P-value < 0.05). Terms were separated in upper and lower boxes because of scale.



Supplementary Figure 3 3. Body weight, bone strength, and IBD antibody for broiler chickens.

Leaf 0.3 and Leaf 0.5 showed higher body weight (P-value: 0.001 and 0.052, respectively)

compared to the control group

This chapter will be published elsewhere
as a partial fulfillment of Sohyun Bang's M.Sc program.

Chapter 4. The overlap between lncRNA and mRNA causes misleading transcripts quantification: A comprehensive evaluation of quantification for RNA-Seq data

4.1 Abstract

Of the immense amount of biological data is arising in this ‘omic’ era, the transcriptome expression pattern is one of the most practical keys to deciphering the genetic information in the whole genome scale. For precise quantification, more controlled and species-specific studies are standard for researchers. One of the limitations in precise quantification is caused by the overlap regions between transcripts, which cause incorrect abundance estimation of those transcripts. From a pool of noncoding RNAs, the long noncoding RNAs (lncRNAs) are known to be overlapped with mRNAs; however, they are generally ignored in previous transcriptome analysis. To illustrate the consequences of disregarding the overlap problem, we examined the error caused by mRNA and lncRNA from all possible annotation databases in 22 species, highlighting its effect on quantification. Here, we defined the overlap region between mRNA and lncRNA for 22 species in all available annotation databases. In some combinations of the database, over half of the protein-coding gene regions are overlapped with the lncRNA loci. The degree of error was estimated according to the use of annotations: quantification using mRNA-only and mRNA-lncRNA annotation. In the mRNA-only quantification, the lncRNA-error rate, which is the ratio of falsely quantified lncRNA-derived reads was high. On the other hand, the mRNA-error rate, which is the ratio of falsely quantified mRNA-derived reads was higher in the mRNA-lncRNA quantification. These error rates increase with higher ratio of overlapped transcripts or longer overlap length. Surprisingly, the increase in the read length or fragment length did not reduce the error rate; instead, it worsened the total error using the most cited tools. As a temporary measure, we suggest some aligners, quantifiers, library construction, and third generation

sequencing on minimizing the errors. Without a doubt, the current study will contribute to all mRNA expression studies by improving the accuracy in actual transcriptome measurement.

4.2 Introduction

The transcriptomic examination of diseases and biological mechanisms have been actively performed on a variety of platforms. Especially, the gene expression levels give one of the more direct quantitative inference of the relationship between phenotype and biological markers. It is currently considered as the most practical and representative biological information (Schena, Shalon et al. 1995, Wang, Gerstein et al. 2009). The advance in sequencing technology such as CAGE and RNA-seq enabled genome-scale investigation of the genes (Kawaji, Lizio et al. 2014). Under the circumstances, accurate quantification of gene expression level is a common goal for researchers; as a result, gene annotation databases have been actively cumulating (Hubbard, Barker et al. 2002, Ozsolak and Milos 2011). Due to a lack of information, the role of RNA was limited as a medium between DNA and Protein, and therefore the gene annotations were focused solely on the protein-coding mRNAs (Eddy 2001, Karolchik 2003, Mercer, Dinger et al. 2009). While the past works focused on the protein-coding genes, current works handle extensions from those genes more frequently. As the non-coding RNAs such as miRNA can regulate gene expression levels, research interest in ncRNA annotation has been increasing (Esteller 2011, Pruitt, Tatusova et al. 2012). Among the ncRNAs, long non-coding RNA (lncRNA) is currently on the top of the list for geneticists (Kung, Colognori et al. 2013).

The lncRNAs, which are non-protein-coding RNAs over 200bp in length, have been recently discovered, and are expected to be widely spread across the genome (Mercer, Dinger et al. 2009, Kung, Colognori et al. 2013). Some reports have elucidated the importance of their functional regulation of transcription and splicing (Mercer, Dinger et al. 2009, Guil and Esteller 2012). While investigating the

characteristics of these lncRNAs, some reports suggest they are overlapped, in genomic position, with the mRNAs they are regulating (Kung, Colognori et al. 2013, Pelechano and Steinmetz 2013). The increase in lncRNA annotation undoubtedly extends the boundary of biological interpretation, yet gives rise to some problems (Kapranov, Willingham et al. 2007). In most transcriptome analyses, the libraries were made with polyA selection to capture the polyA tails of the mRNAs; however, recent works announced that some lncRNAs also have polyA tails also (Mattick and Makunin 2006, Mortazavi, Williams et al. 2008). Hence, the previous works that were supposed only to quantify the mRNA expression, may include lncRNA expression within the library (Kapranov, Cheng et al. 2007). Such symbiotic existence of lncRNA intensifies annotation-based transcriptome analysis error in two broad categories: (i) the lncRNA expression exists, but is not measured at all; (ii) the lncRNA expression is treated as mRNA expression. The first error can be eventually fixed through lncRNA annotation expansion. As for the second type, the error intensifies with the increase in lncRNA annotation database size due to the overlap problem. Hence, an in-depth study on the error rate of expression level quantification of the overlap region is imperative. Although there were comprehensive studies of overlap region in certain species, they lack focus on the lncRNA-mRNA overlap and between-species comparisons of those regions (Makalowska, Lin et al. 2005, Sanna, Li et al. 2008). Therefore, a comprehensive summary of the errors caused by the mRNA-lncRNA overlap has to be established in a timely manner.

Unlike the aforementioned overlap problem, similar ambiguous reads problems— isoform problem (Schuierer and Roma 2016) and mRNA-mRNA overlap problem (Sun, Yang et al. 2015)) —have been investigated, and in vitro or in silico methods were proposed to disburden those problems (Katayama, Tomaru et al. 2005,

Levin, Yassour et al. 2010). Under in vitro conditions, a strand-specific protocol during cDNA synthesis is a partial solution for the overlap in the complementary strand (Sigurgeirsson, Emanuelsson et al. 2014). Also, there are methods like re-designing the microarray probes, to consider the expression levels in the overlap region (Yelin, Dahary et al. 2003). Under in silico conditions, the overlapped region problems are handled through algorithms in the quantification steps that consider repetitive features induced by alternative splicing and overlap region (Sun, Yang et al. 2015, Schuierer and Roma 2016). Although there were some attempts on mitigating the overlap problem, a systematic investigation of such errors caused by mRNA and lncRNA overlap needs to be conducted in several species for developing an improved version from the current quantification algorithm.

Today, five popular databases—NONCODE, PLAR, ALDB, GENCODE, and LNCipedia—lead the lncRNA annotation area, which respectively provides information for 17, 16, 3, 2, and 1 species (Derrien, Johnson et al. 2012, Hezroni, Koppstein et al. 2015, Li, Zhang et al. 2015, Volders, Verheggen et al. 2015, Zhao, Li et al. 2016). Furthermore, the two representative databases for mRNA annoatation—Ensembl and Refseq—are also going through expansions to include the lncRNA annotations (Pruitt, Tatusova et al. 2012, Yates, Akanni et al. 2016). In this study, all of the available annotation databases were employed to define the overlap region in 22 species. We observed how much of the so-called ‘lncRNA error rate’ and ‘mRNA error rate’ emerge from the overlap regions. To solve this problem, we simulated various sequencing systems such as read length and fragment length. Furthermore, we evaluated aligners and quantifiers regarding these error rates.

4.3 Materials and Methods

4.3.1 Collecting lncRNA annotations from 7 databases

All accessible lncRNA annotation was employed from seven databases: NONCODE, PLAR, ALDB, LNCipedia, GENCODE, Ensembl and RefSeq (Borodina, Adjaye et al. 2011, Bu, Yu et al. 2011, Derrien, Johnson et al. 2012, Pruitt, Tatusova et al. 2012, Hezroni, Koppstein et al. 2015, Li, Zhang et al. 2015, Volders, Verheggen et al. 2015). The NONCODE provides annotation of 16 species, which are collected from literature published and several public databases. The PLAR annotated lncRNAs of 17 species with their in-house pipeline. The ALDB annotated lncRNAs of three domestic animals with a focus on long intergenic noncoding RNAs. The LNCipedia annotated lncRNAs of human, and lastly, GENCODE provides human and mouse annotation. Among 39 species, seven species have been annotated in multiple databases: Human in four databases, Mouse and Chicken in three databases, Zebrafish, Cow, Opossum and Rhesus in two databases. A total of 27 species were annotated in at least lncRNA database. In addition, Ensembl and RefSeq provide non-coding RNA annotation, which is defined as ELA and RLA, respectively. The ELA and RLA were extracted from the original annotation set. ELA annotated lncRNAs of 9 species, and RLA annotated 16 species among 27 species which was annotated in five lncRNA databases.

4.3.2 Collecting mRNA annotations from Ensembl and RefSeq

Ensembl and RefSeq mRNA annotation are widely used in biological research (Levin, Yassour et al. 2010). Based on the assembly version that is used in lncRNA

annotation, release for mRNA databases such as Ensembl and RefSeq was determined. Among 27 species, some species with only lncRNA DB were excluded. In RefSeq, Stickleback, Elephant shark and Ferret were excluded as their assembly versions were not matched with those of lncRNA DBs. Also, in Ensembl, Rhesus monkey, Stickleback, Nile tilapia and Elephant shark were not considered due to different assembly version with lncRNA DBs. Both of Ensembl and RefSeq (from UCSC) do not provide reference and mRNA annotation for Ferret and Spotted gar, which is annotated in PLAR. By excluding these species, 22 species in total were both annotated in mRNA database and lncRNA databases with matching assembly versions. We obtained Ensembl annotation file from Ensembl FTP site and RefSeq Gene annotation file from UCSC Table Browser (Robinson and Oshlack 2010). RefSeq provides BED format and GTF format in UCSC, and Ensembl provides GTF format.

4.3.3 Statistics from annotated mRNAs and lncRNAs in 22 species

For statistics by transcript level, the BED format is employed to consider the position of transcripts. In some annotation sources—Ensembl, ALDB, and GENCODE—that do not provide BED format, we converted the GTF files to BED format. By each annotation, the number of transcripts and the mean of transcript lengths were summarized along with their standard error of the mean (SEM). Moreover, the mean of three more characteristics—exons per one transcript, length of exons, and length of introns in transcript—were calculated with their SEM. To present the difference of databases, we calculated the average of the mean number of exons and mean length some species that have more than two databases.

In order to consider the genomic position of exons, we employed the GTF format. In some annotation sources that do not provide the GTF format, GTF format annotations were generated by converting provided BED formatted files (transcript unit). The GTF format contains the same position of exons to present units of gene and transcript. To only consider the exonic positions, repetitive exons were excluded. By each annotation, the number of exons and mean of exon lengths were respectively calculated with their SEM. To present difference of databases, total means of the two measures—the mean number of exons and mean length—were calculated in the species that have more than two databases.

4.3.4 Defining overlapped regions between mRNA and lncRNA

Overlapped regions between mRNA and lncRNA exons are found when either start or end position of exons in mRNA annotations are included in exon positions in lncRNA annotations, and vice versa. The overlaps are detected for each chromosome (scaffold was not considered) in all species. Using the positions of overlap pairs, we sorted different-strand overlaps and same-strand overlaps. We calculated overlap length only considering the overlapped exon length.

4.3.5 Calculating ratio (D)

The Ratio of overlapped mRNA exons to lncRNA exons determines the overlap shape. The number of mRNA exons was divided by that of lncRNA exons and the ratio presents that how many mRNA exons make a pair with lncRNA exons averagely. By dividing the number of overlapped lncRNA exons from the number of overlapped mRNA exons, a ratio (D) that represents an average shape of overlap can be achieved.

When the ratio is over 1, it means that more than one mRNA exon are overlapped with one lncRNA exon on average. Log scale was used here to demonstrate that the ratio smaller than one is equally as influential as the ratio higher than one. The ratio of log scale is calculated in 60 combinations of databases.

4.3.6 Read generation using flux simulator

Flux simulator v1.2.1 was used to generate reads, extracted from all combinations of mRNA and lncRNA annotations (Love, Huber et al. 2014). Limitations of simulated data include an imperfect ability to model sample preparation protocols and the distribution of reads across the genome. The Flux Simulator tries to overcome these limitations by the use of parameterized models that reproduce the common sources of systematic bias in each of the main steps in sample preparation. To achieve uniform expression of all genes (mRNAs and lncRNAs), while considering the transcript length, the following options are used: FRAG_UR_ETA; 500 and PCR_PROBABILITY; 0.05.

In the error profiling in whole transcripts, we generated 80 million from all annotated transcripts, which have approximately 300 reads per transcript. For other simulations which sampled transcripts (such as sampling overlap pairs with 100-200bp overlap legnth), 300 reads were also generated per transcript. To control the number of reads, read length, and fragment length, we ran the simulator with the following options using PAR file; 'READ_NUMBER', 'READ_LENGTH.', and 'FRAG_UR_ETA '. Additional parameters were supplied in a parameter file as outlined in the Flux Simulator manual available at: http://sammeth.net/confluence/display/SIM/Home (accessed 24 July 2018).

4.3.7 Common simulation setting

Produced reads were cleaned by Cutadapt (v 1.14) (Martin 2011). After trimming, trimmed reads were aligned to the genome using the HISAT2 alignment software (ver 2.0.0) (Kim, Langmead et al. 2015). The genome assembly—GRCh38, GRCm38, Release.6.plus.ISO1.MT—were employed. We quantified the mapped reads using featrueCounts (Liao, Smyth et al. 2013) with the annotation file (.GTF) for mRNA or mRNA-lncRNA. For mRNA-lncRNA quantification, the GTF file for mRNA and lncRNA are merged.

4.3.8 Sampling transcripts for simulation.

For the simulation of overlap errors according to the overlap length, the overlap-length were binned every hundred bp from 1 to 1000. We randomly sampled the mRNA-lncRNA paired reads each and repeated this steps three times. The number of selected sample was 90% of the minimum number from 10 bins.

For the simulation of mRNA-lncRNA's directional property, we sampled the overlapped transcripts in different strand overlap or same strand overlap. The ratio of different strand (or same strand) overlap was binned from 10% to 90%. The ratio of different strand overlap and same strand overlap were respectively represented by D and S. The selected transcripts in the bin is also random and repeated three times.

4.3.9 Regulating feature-overlap option in FeatureCounts

Feature-read overlap option was employed: mRNA-lncRNA overlap length of 100-200 and 900-1000 samples were the basis for the feature-read overlap option (--minOverlap). The min overlap option was regulated from zero to the length of the read.

4.3.10 Strand-specific RNA-seq data

FLUX did not provide to generate reads with the strand-specific library. To make strand-specific reads using FLUX, we produced reads and then selected reads which are from the original strand of the transcript. This is similar with dUTP, NSR, NNSR strand-specific library construction (Borodina, Adjaye et al. 2011). Reads were also aligned with HISAT using the '--rna-strandness RF' option and quantified with featureCounts using '-s 2' option.

4.3.11 Aligner simulation

In an attempt to illustrate our concerns in the current RNA-seq protocols, we evaluated the five most popular and concurrent aligners: HISAT2, STAR, OLEGO, TOPHAT2, and SUBREAD (Mortazavi, Williams et al. 2008, Robinson, McCarthy et al. 2010, Martin 2011, Bolger, Lohse et al. 2014, Ayers, Lambeth et al. 2015). As aforementioned, we carried on this analysis with the human GRCh38 RefSeq and NONCODE database combination, since it gave the lowest error rate in our previous step. The default parameters were employed with the exceptions on strand-specificity and multi-threads (8 threads) options if necessary: The HISAT2 required '-p 8 --rna-strandness RF'; for STAR, '--runThreadN 8'; OLEGO, '-t 8 and --strand-mode'; TOPHAT2, '-p 8 --library-type fr-firststrand', and lastly, the SUBREAD required '-

T 8 -S fr'. Note that the STAR does not require a strand-specificity option as it automatically detects the library type. The number of cores is fixed at 8 to compare the running time between the aligners. By fixing featureCounts as the quantifier for all aligners, we compared the results for error rates.

4.3.12 Quantifier simulation

Following the aligner, we tested two of each aligner-independent (HTSeq and FeatureCounts) and –dependent quantifiers (eXpress and Salmon) to display our concerns in the quantification step (Griebel, Zacher et al. 2012, Liao, Smyth et al. 2013, Roberts and Pachter 2013, Anders, Pyl et al. 2015). The aligner-independent quantifiers used Hisat2 for mapping, bowtie2 was used for eXpress, and Salmon's quasi-mapping was used for Salmon's optimal results. Since eXpress and Salmon uses transcript-based reference (eXpress style multiFasta), HISAT2's split-read mapping provided no advantage; bowtie2 and Salmon's mapper gave better results, and more detailed mapping information could be extracted when the recommended aligners are used. Default parameters were used in the analysis, and some options that do not alter the results were added (i.e., program log and mapping information files). For HTSeq, '-r pos -s reverse' options were included and as for featureCounts, '-T 10 -s 2 -R –p' options were used. Detailed descriptions are in the user guides, but the included parameters are for input-specificity, thread, or read-report options. As for the aligner-dependent duo, '--no-update-check --rf-stranded --output-align-prob' was used for eXpress, and '-p 10 -l ISR -s' for Salmon. As mentioned above, the options are input-specific, and most parameters are set as default.

4.3.13 Simulation for long read from third generation sequencing

Reads were generated from the full length of transcripts. The reads with error rate from third-generation sequencing such as Nanopore is estimated to be higher (about 10%) compared to that of second-generation sequencing such as Illumina (less than 0.1%) (Aken, Achuthan et al. 2016, Tyner, Barber et al. 2016). Thus, we generated long reads with 10% base error rate by randomly creating a false nucleotide base. For alignment and quantification, bowtie2 and eXress were used, respectively.

4.4 Results

4.4.1 Baseline characteristics of the collected mRNA and lncRNA annotations in 22 species

We collected all possible pairs between mRNA and lncRNA annotations resulting in 50 mRNA and 59 lncRNA annotations across 22 species, for a systematical investigation of the overlap pattern in each species (Supplementary Figure 4.1). In species other than Opossum, the number of annotation for protein-encoding transcript was averagely 2.58 times larger whereas the number of protein-coding exons was 5.76 times larger in all species (Supplementary Figure 4.2). In C.Elegans, Macaque, Marmoset, and Zebrafish species, a larger number of annotation differences than other species was found between mRNA and lncRNA (10.58-fold and 27-fold larger at the transcript and exon levels, respectively). This is due to the the less annotated lncRNAs in certain species, and reveals that lncRNA annotation is still unstable in those species (Supplementary Figure 4.2). This instability has been confirmed once again by examining the variability of several different versions of the lncRNA annotations (Supplementary Figure 4.3). Based on the dispersion of mRNA annotations, the dispersion of lncRNA annotations was higher in species except Dog and Opossum in terms of the number of annotations and the length of transcripts. This suggests that heterogeneity issue could arise in the quantification of gene expression levels according to the selection of a combination annotations exhibiting different characteristics within certain species. Assuming no overlaps between transcripts, an average of 12758.81 additional transcripts were added over 22 species, with an average of 28691.69 exons added at the exon level, if lncRNA annotations were additionally considered in the transcriptome analysis. Despite the fixed size of the

genome, the increased number of annotations due to the addition of lncRNA information indicates that the overlap between mRNA and lncRNA will be intensified.

4.4.2 Protean overlap pattern between mRNA and lncRNA

We hypothesized that these various annotation combinations would represent various overlapping patterns and that this pattern had a direct impact on the quantifying gene expression levels. As a first step, we investigated the overlap between comparable mRNA and lncRNA annotations. A total of 119 pairs of mRNA and lncRNA annotations were available in collected data (Supplementary Table 4.1). After filtering pairs containing unreliable annotations (< 6.6k mRNAs and <1k lncRNAs), a total of 72 mRNA-lncRNA combinations remained across the 20 species (Supplementary file 1 - Table S4 and S5). On average, 20,063 mRNA-lncRNA overlapped transcripts were found across all combinations and 19.1% of the lncRNA and 11.24% of the mRNA annotations overlapped with each other (Figure 4.1a and 4.1b). Based on the mRNA annotation, Fruitfly (33.01%), Rat (21.37%), and Human (17.52%) species showed the highest overlap, due to lncRNAs derived from NONCODE. The average overlapped mRNA ratio of each lncRNA annotation was higher in order of LNCipedia (47.65%), NONCODE (23.93%), RLA (9.6%), GENCODE (9.3%), PLAR (5.85%), ELA (1.07%), ALDB (0.2%). Based on the mRNA annotation, we found that the combination with Refseq (15.25%) showed a higher overlap ratio than the combination with Ensembl (9.12%). These results indicate that the overlapping degree is significantly different depending on the combination of mRNA-lncRNA annotation and can be exacerbated in certain species. In addition, this showed that the degree of overlap can vary greatly depending on the selection of lncRNA annotation rather than mRNA annotation.

This phenomenon was also observed when examining the structural characteristics of the overlap region on the genome. Investigating at the characteristics of the identified overlap pairs based on the strand information of the

associated genome, the stranded ratio difference between mRNA annotations was relatively less varied than the stranded ratio difference between lncRNA annotations (Figure 4.1c). Overlapped transcripts with RLA (Avg. 59.51%) and LNCipedia (Avg. 73%) had a high ratio of same stranded overlap across all species, while PLAR (Avg. 7.56%) and GENCODE (Avg. 12.55%) showed low ratios of same stranded overlap in most species (Supplementary Figure 4.4). This result indicates that the overlap of mRNA-lncRNAs in the same and different strands is present in all species and that the ratio of the same or different stranded overlap varies dramatically depending on the mRNA-lncRNA combination. This result is important because the mRNA-lncRNA overlap located in the same strand cannot be resolved in current RNA-sequencing based on short read, whereas other stranded overlap can be resolved by the strand-specific library technology.

In the similar vein, we suspected that variable factors of RNA-sequencing (i.e. single or paired-end libraries, fragment size) could affect the classification possibility of the mRNA-lncRNA overlap according to their overlap length and structure. The overlap length for all mRNA-lncRNA combinations at transcript unit was 552.47bp on average. Considering that the most widely used fragment size in current RNA-sequencing is 500bp, this shows that over half of the mRNA-lncRNA overlapped pairs exceeds the expected fragment size. We also examined the exonic structure of these overlapping regions within each transcript because performance of the split-read alignment method used in most common quantifiers for considering alternative splicing could be highly affected in three or more exons of both mRNA and lncRNA. Of the entire transcripts with the mRNA-lncRNA overlap we identified, average 44.62% of mRNAs and 59.34% of lncRNAs show overlapping region that span three or more exons. The relative ratio of overlapped exons of mRNA and lncRNA was

examined and it was found that overlap region was distributed over a relatively large number of lncRNA exons (Figure 4.1d). In all of the various aspects we investigated, the variation among lncRNA annotations was larger than that of mRNA annotation, indicating that lncRNA annotation is still unstable.

The number of overlapped pairs had positive correlations with the density of annotated transcripts (Supplementary Figure 4.5). Most of the species display a positive correlation between the number of overlapped pair and the transcript density. However, the non-primates—A.Thaliana, C.elegans, and Fruitfly—show a low number of overlapped pairs compared to their transcript density. This indicates that the transcript density and the number of overlapped pairs is distinctive between primates and non-primates.

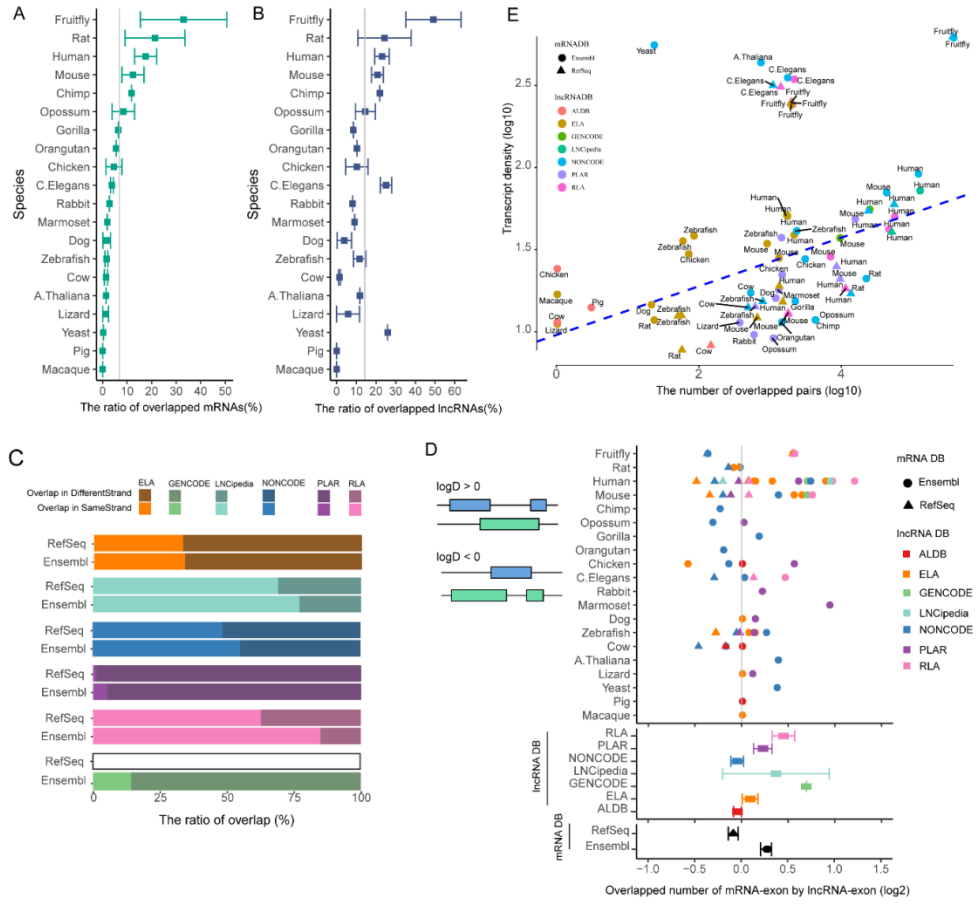


Figure 4.1. Characteristics of annotated transcripts and overlap region.

A) Dot plot show the mean of ratio of overlapped mRNAs between annotation databases. B) Dot plot show the mean of ratio of overlapped lncRNAs between annotation databases. C) Relationship between transcript density and the number of overlap pairs with log scale. The shape of dot represents mRNA annotation databases while the color represents lncRNA annotation databases. D) The Percentage of overlap direction. Darker color (Brighter color) shows the mRNA and lncRNA overlap in different strand (same strand). E) By using the D value from above, the ratio of overlapped mRNA exons to overlapped lncRNA exons are plotted across the species. Since the ratio is log transformed, D=1 is log ratio=0, which is represented with a yellow dotted line.

4.4.3 Error profiling when all transcripts (mRNA&lncRNA) are expressed

In our simulation, we evaluated the mRNA expression quantification error, under the hypothesis that diverse lncRNAs are actually expressed. To do so, we let mRNA and lncRNA be expressed in all database combinations and measured the gene expressions in two ways about what annotations are used, the mRNA-only and mRNA-lncRNA quantification. The mRNA-only pipeline uses only mRNA annotation for quantification while mRNA-lncRNA quantification use both annotations for quantification (Table 4.1). Two tables of confusion were constructed based on the origin and quantified RNA, which are actual class and predicted class, respectively. Here, the mRNA-error rate is defined as the number of false negatives over all mRNA-derived reads, while the lncRNA-error –rate is defined as the number of false positives over all lncRNA-derived reads. To present the degree of error in mRNA-reads and lncRNA-reads, lncRNA-reads derived error rate (lncRNA-error), mRNA-reads derived error rate (mRNA-error) and Total Error rate (TER) were investigated. For the three most practically studied species—Human, mouse, and fruitfly—error assessment was performed for all mRNA-lncRNA database combinations (Figure 4.2 and Supplementary Figure 4.6).

The lncRNA-error rate was high only in mRNA quantification. This is caused by the lncRNA-reads that are aligned to the overlap position and thus quantified as mRNA with absence of lncRNA annotation in quantification. lncRNA-error rate stays at zero in mRNA-lncRNA quantification because lncRNAs-reads in overlap region are rarely counted to mRNA when both mRNA and lncRNA annotations were used. On the other hand, the mRNA-error was higher in mRNA-lncRNA quantification compared to mRNA quantification. Overlapped features such as mRNA-mRNA or

mRNA-lncRNA produce mRNA-reads that are not quantified as mRNAs. The probability of unquantified mRNAs is higher in mRNA-lncRNA quantification. Thus, error caused by mRNA quantification is mainly lncRNA-error, while error in mRNA-lncRNA quantification is mRNA-error. Herein, we can refer to the lncRNA-error rate as the error from mRNA quantification due to lncRNA, while referring to the mRNA-error rate as the error from mRNA-lncRNA quantification due to mRNA. The total error rate (TER) is highly affected by mRNA-error because the number of conditions in mRNA-error is significantly greater than that in lncRNA-error. Thus, the TER was generally higher in mRNA-lncRNA quantification except for some annotation combinations in fruitfly.

The defined errors were different according to the combinations of databases. In the case of Human, the lncRNA-error rate was over 50% in some database combinations including RLA lncRNA annotation—the Ensembl&RLA and RefSeq&RLA. These combinations also had high mRNA-error rate. These affect highest TER in the RefSeq-RLA (13%) in mRNA quantification, and Ensembl-RLA (9%) in mRNA-lncRNA quantification. On the other hand, the ELA-including combinations show low lncRNA-error and mRNA-error rates that are under 0.2%. As a result, the RefSeq-ELA combination has the lowest TER of 0.1% in both quantifications (mRNA-only and mRNA-lncRNA annotation). We focused on the TER differences based on the change of mRNA and lncRNA databases; the change of database affects the TER respectively by 0.3% and 1% for mRNA and lncRNA on average. We therefore deduced that annotation, especially the lncRNA annotation, has an influence on the error rates since the characteristics of each DB are different.

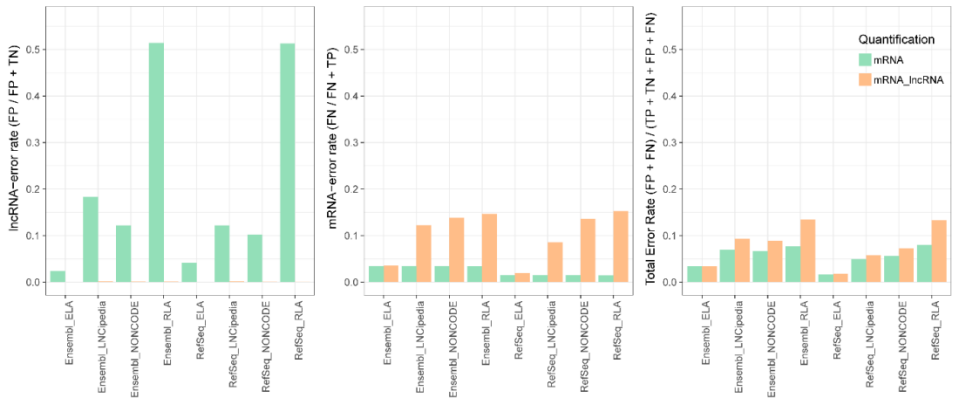


Figure 4.2. Simulation for whole transcripts in combinations of human databases.

This shows mRNA error, lncRNA error, and total error rate in all database combinations for Human. The mRNA-only and mRNA-lncRNA quantification are respectively colored in green and orange.

Table 4.1. Contingency table for reads in two quantification pipeline

		Quantification in mRNA-only quantification pipeline	
		mRNA	Not mRNA (e.g. No feature, ambiguous)
Origin (Derived-)	mRNA	True Positive	False Negative (mRNA-error)
	lncRNA	False Positive (lncRNA-error)	True Negative

		Quantification in mRNA-lncRNA quantification pipeline	
		mRNA	Not mRNA (e.g. No feature, ambiguous, **lncRNA**)
Origin (Derived-)	mRNA	True Positive	False Negative (mRNA-error)
	lncRNA	False Positive (lncRNA-error)	True Negative

4.4.4 Error profiling according to overlap length between mRNA and lncRNA

The error related to the overlap length was estimated by simulating transcript length from 1-100bp to 900-1000bp at a 100bp interval (Figure 4.3 and Supplementary Figure 4.12). All errors tend to increase as the overlap length increased; overlap length from 1-100bp to 900-1000bp results in an average increase of 50% (10% to 60%) in the lncRNA error rate. With the increase in the percentage of overlapped transcripts, which is compatible to the overlap length, the mRNA-error rate also increases in an average of 19% (0% to 19%). Compared to the mRNA error rate, lncRNA error rate had higher increase. In summary, overlap length was associated with all errors, with an exceptionally high increase in lncRNA-error rate.

The degrees of error were different between species. Compared to the Human and Mouse, lncRNA-error in Fruitfly had the irregularity of relationship with overlap length and higher error variance. This irregularity is expected to be caused by transcript lengths in Fruitfly, which had higher variance between replicates in overall sampling intervals (Supplementary Figure 4.7). This dissimilarity was associated with the mRNA and lncRNA length. In both simulations, Mouse had longer lncRNA length compared to mRNA length while Human had similar length of mRNA and lncRNA. Therefore, there are features other than overlap number and overlap lengths (i.e. transcript length) that can cause a difference in the between-species error rates.

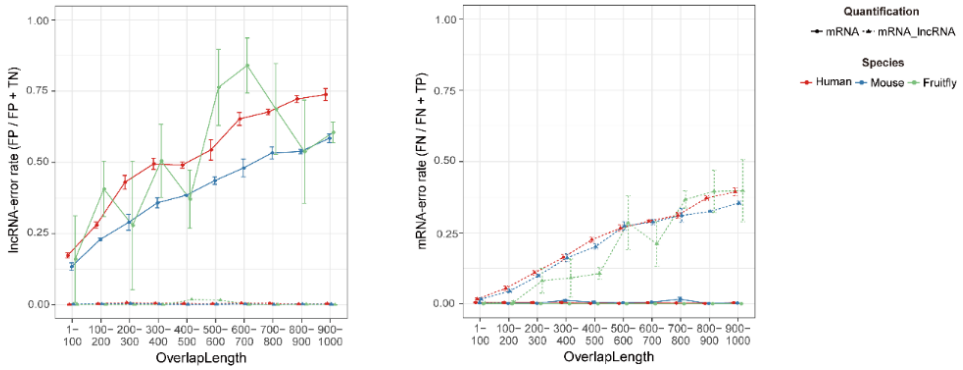


Figure 4.3. Simulation for error according to the number of overlap and overlap length.

The three different species, Human, Mouse, and Fruitfly, are color-coded (red, blue, and green). The solid and dashed lines correspond to the mRNA-only and lncRNA-mRNA quantification, respectively. a) error according to the percentage of overlapped mRNA. b) error according to the overlap length (bp).

4.4.5 Error profiling with read length and fragment length

Read length and fragment length are components of sequencing systems. As they are assumed to be associated with overlap error, we performed a simulation of error according to read and fragment length. We expected that long read and fragmentation can reduce error when they cover the overlap region. To investigate the relationship between overlap length and these components of sequencing system, we simulated error according to read length or fragment length in selected combinations of DB for Human. Surprisingly, the longer the read-length or fragment-length, the higher the error rate was.

To simulate the impact of read length on error, we generated various read length from 25bp to 250bp with same libraries (Figure 4.4a and Supplementary Figure 4.8). In most overlap length, the error rate was higher in the longer reads (Figure 4.4a). In overlap length 900-1000bp, the TER for mRNA quantification was 23% in 250bp, which is 3% higher than that in 25bp. The mRNA-error rate was slightly higher in the longest reads (250bp) compared to the shortest (25bp). The difference between the shortest and longest reads was highest in mRNA error rate, with 10% difference. The higher error rate in longer reads is due to the quantification algorithm, which maps a read on a feature even with a single base pair inclusion. Simply put, a longer read has a higher chance of being included as a feature under such settings.

To solve the defective quantification of longer reads, we regulated options for read-feature length in FeatureCounts, which determine whether the read is counted as feature based on read-feature overlap length (Supplementary Figure 4.9). Although lncRNA-error rate decreased in longer read-feature overlap options, the mRNA-error

rate also increases with increase of length for feature-read ovrlap option. In case the overlap length is 100-200bp, the feature-read overlap option of 0.4 to 0.5 seems to reduce the TER of the longer reads. With minor differences, the TER of overlap length 900-1000bp seems to be unaffected; this may be due to the higher lncRNA-error rate in longer overlapped criteria. Considering the distribution of overall overlap lengths, the feature-read overlap option of 0.4-0.5 is plausible solution for a considerable reduction of the lncRNA-error rate in a whole genome analysis.

For the simulation of error according to the fragment length, samples various fragment length from 400bp to 900bp with same read length was generated (Figure 4.4b and Supplementary Figure 4.10). Overall tendency is that the increase in fragment length results an increase in the lncRNA-error rate. The same property behind the impact of read length on error, applies to this analysis. Compared to the simulation for read length, simulation for fragment length showed a higher difference of lncRNA-error rate between the shortest and longest fragment, especially in some overlap bins. In the 100-200 to 300-400 overlap bins, lncRNA-error rate increased with the fragment size while from 300-400 to 900-1000 displays a decreasing pattern. This difference is also associated with different overlap shapes, which affect lncRNA error rate. There are three overlap shapes that can be considered: 1) 1:1 overlap, 2) $D>0$, and 3)$D<0$, the selected RefSeq-NONCODE combinations of human is a $D<0$ shape where there are more than one lncRNA per mRNA. In general, the bins with higher lncRNA error rate will have a D ratio greater than that of the low lncRNA error rate bins. The difference of mRNA-error rate between longest and shortest fragments are minute; and the fragment size and mRNA-error rate does not have a clear linear relationship between overlap length and fragment length.

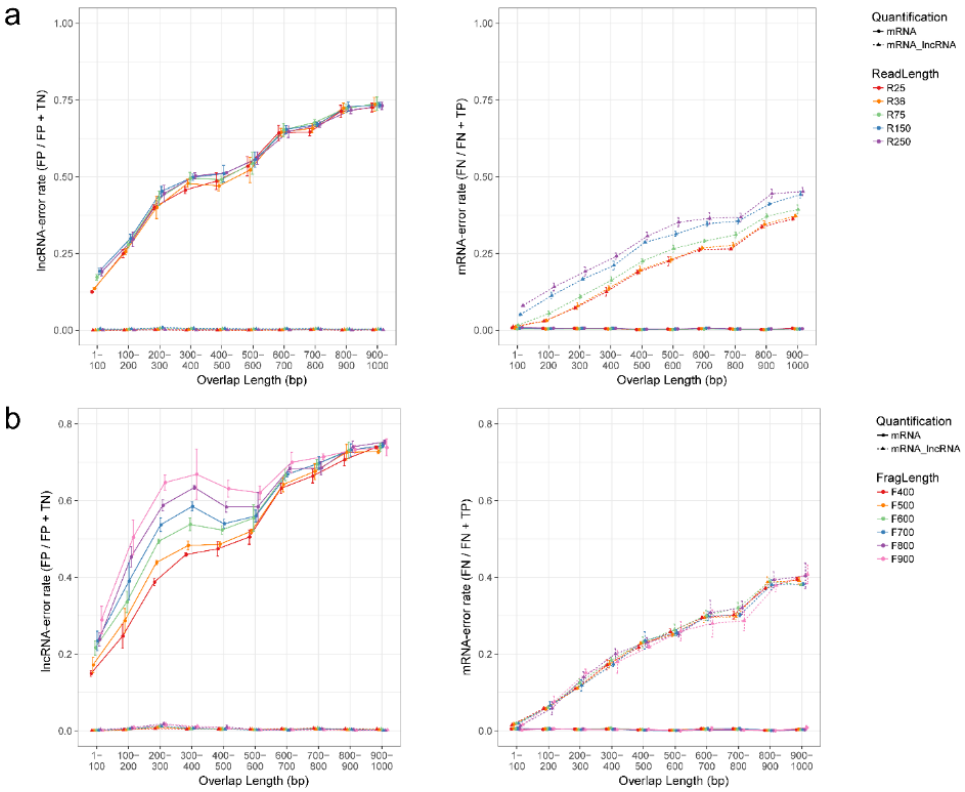


Figure 4.4. Simulation for error according to read length and fragment length.

The solid and dashed lines correspond to the mRNA-only and lncRNA-mRNA quantification, respectively. a) the relationship between read length and overlap length in the error rate. The different read length from 25 to 250 bp were color-coded. b) the relationship between fragment length and overlap length in the error rate. The different fragment length from 400 to 900 bp were color-coded.

4.4.6 Error-reducing pipeline using available resources

We focused on the available tools that may decrease the overlap error in RNA-seq analyses. The use of strand-specific over non-strand-specific library, change of aligners and quantifiers, and finally, investigation of error in long reads are illustrated. We employed only the human samples for these comparisons under RefSeq-NONCODE combination.

The errors due to the strand-specific library and overlapped mRNA-lncRNA's directional property—same-strand overlap and different-strand overlap—has been evaluated. We analyzed how much of the different-strand overlap can be solved with the strand-specific library; the ratio of the direction of overlap and pipeline-specific results have been considered. In the mRNA quantification, strand-specific pipeline proved to be better in terms of total error and lncRNA error rates (Figure 4.5a and Supplementary Figure 4.11). The more different-strand overlap is prevalent (greater D), the lesser the error rates are in the strand-specific pipeline. Thus, the strand-specific pipeline effectivley reduced errors caused by the same-strand overlap. To check the error performance of the strand specific pipeline in the whole transcriptome level, we simulated all transcripts annotated in RefSeq and NONCODE using strand specific pipelines (Figure 4.5b). As a result, the lncRNA-error rate shows a significant decrease, almost by half, and the mRNA-error rate and TER decreased slightly.

To analyze if there is an optimal error-reducing algorithm among the aligners, we evaluated five of the most popular aligners—Tophat2, Olego, Subread, Hisat2, and STAR—on their error performances under strand-specific conditions (Mortazavi, Williams et al. 2008, Robinson, McCarthy et al. 2010, Martin 2011, Bolger, Lohse et

al. 2014, Ayers, Lambeth et al. 2015). In terms of TER, the Olego and Subread displayed higher error rates compared to the other three (Figure 4.5c and Supplementary Figure 4.12); the difference in TER is caused by the mRNA-error rate, since the lncRNA-error rates show no significant differences. Interestingly, the Olego and Subread, which have higher mRNA error rate, display higher mapping rates; Olego (65% of total reads mapped) and Subread (68%) had higher mapping rates than the other three (~60%). While Tophat2, Hisat2 and STAR show solid performances, when error rate is the only factor, but Hisat2 and STAR are top of the line when running time is also considered. Also, STAR and Hisat2 are the most frequently cited standalone aligner tool for RNAseq studies in the latest publications.

Following the aligner study, we tested out two of each aligner-independent (HTSeq and FeatureCounts) and –dependent quantifiers (eXpress and Salmon) (Griebel, Zacher et al. 2012, Liao, Smyth et al. 2013, Roberts and Pachter 2013, Anders, Pyl et al. 2015). For the aligner-independent quantifiers, we carried on the alignments with Hisat2, which had the best performance in the aligner simulation. Here, the featureCounts dominantly outperformed HTSeq in all error rates (Figure 4.5c and Supplementary Figure 4.13). For the aligner-independent quantifiers, we chose the flagship program, eXpress (with bowtie2), and the most concurrent tool, Salmon, for simulation. While HTSeq and featureCounts use gene-level summaries, these two programs use transcript-level suammry of the transcriptome. The two transcript-based aligner-dependent tools both show higher TER in the mRNA quantification pipliene compared to mRNA-lncRNA quantification. This is caused by higher lncRNA-error rate under mRNA quantification. This is unique error pattern because aligner-independent tools had lower TER caused by lower mRNA-error rate in the mRNA quantification over mRNA-lncRNA quantification. When comparing

Salmon with eXperess, eXpress had lower TER in mRNA-lncRNA quantification while Salmon had better performance on mRNA quantification. However, Salmon's lncRNA-error rate is almost 100% under mRNA quantification. The Salmon tries to align much more transcripts than eXpress (84,807,672 vs. 17,842,038 total transcripts) and allows more transcripts to be multimapped, one transcript can be mapped to more than 10 different loci while eXpress doesn't. Such quantification in the mapped regions, may results in the increase of lncRNA-error. Therefore, eXpress had better performance in terms of all error rate than Salmon. In summary, the aligner-dependent quantifiers show better performances in mRNA quantification, while the aligner-independent quantifiers are better under mRNA-lncRNA quantification. We also confirmed that featureCounts (in mRNA) and eXpress (in mRNA-lncRNA) have lowest error rates under each condition.

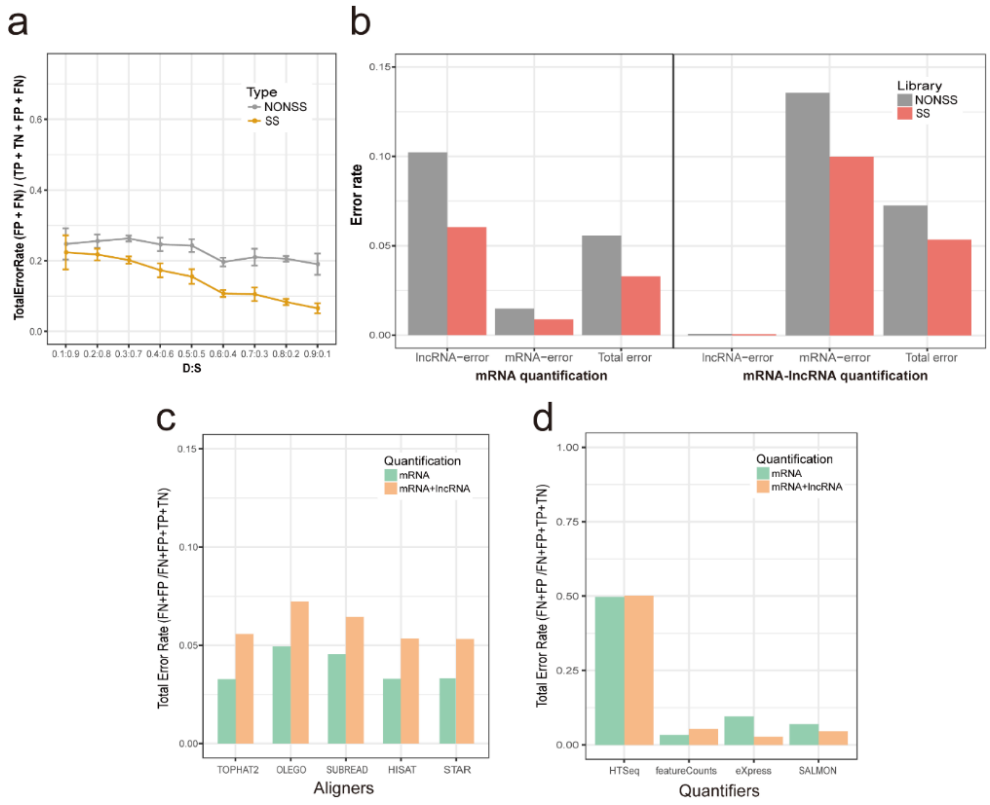


Figure 4.5. Error rates for error-reducing pipelines.

The figure is for human RefSeq-NONCODE combination analyses. a) The error comparison between strand-specific library vs. non-strand-specific library under different D:S ratios (here, D: ratio of different-strand overlap and S: ratio of same-strand overlap); b) The error performance between five different aligners (Tophat2, Olego, Subread, Hisat2, and STAR); c) The error performance between four quantifiers, two aligner-independent (HTSeq and featureCounts) and two aligner-dependent (eXpress and Salmon) quantifiers.

4.4.7 Error profiling in third generation sequencing

From our results, generating strand specific reads and using HISAT & featureCounts (for gene level quantification) or eXpress (for transcript level quantification) have the optimal pipeline to minimize error. To reduce the error even more by using this optimal pipeline, we examined error rate in third generation sequencing. Direct RNA-seq sequencing in Nanopore does not use PCR and produces long reads that cover whole length of transcript (Garalde, Snell et al. 2018). We analyzed the errors under ideal settings, where the whole length of transcript is sequenced. Especially, these long reads are expected to reduce errors in the longer overlapped regions.

If the overlap length is 900-1000bp, using a short read of 150bp induced a high error rate (Figure 4.6). In mRNA quantification, the lncRNA error rate from 150bp read length is 73%. The long reads reduced half of this lncRNA-error rate, decreasing the TER by 2/3. In mRNA-lncRNA quantification, long read decreased TER by 3/4, displaying a much significant decrease in error compared to mRNA quantification. Such dramatic decrease is due to the mRNA-error rate, which has been reduced from 45% to 3%. However, the lncRNA-error rate increased by 10% in the long read. In summary, we observed successful reduction of TER when long read is employed except for increased lncRNA-error rate in mRNA-lncRNA quantification.

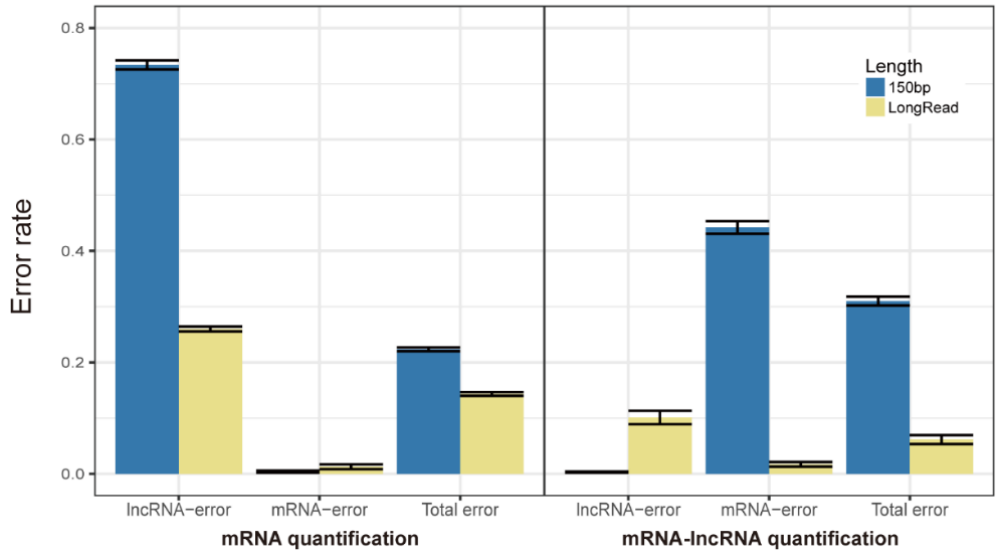


Figure 4.6. Error rate for long read generated from third generation sequencing.

The error rate for the reads from 900-1000bp overlap region. The 150bp read and long read covering a whole transcript are color-coded as blue and yellow respectively.

4.5 Discussion

We confirmed that quantification error is evident where overlapped region exists. However, a majority of the studies did not account for such error. In most studies, the main objective is on the mRNA expression, and therefore only use the mRNA annotation in the quantification step (Ayers, Lambeth et al. 2015). In some cases, where both mRNA and lncRNA expression is considered, two separate quantification pipelines are used to measure each RNA's expression (Gong, Huang et al. 2017). Considering this, we examined the mRNA quantification pipeline, which only considers mRNA annotation. In this pipeline, high lncRNA-error, which is directly related to misquantified mRNA expression level, is observed. In the whole transcriptome error simulation, the lncRNA-error rate in Human could go up to 50%. This lncRNA error rates which caused by overlapped lncRNA reads are crucial to the accurate quantification of the mRNA expression. For instance, when both the mRNA and lncRNA are expressed, inconsideration of lncRNA annotation results in an over quantification of the mRNA expression. In a case where lncRNA expression suppresses the mRNA expression, the addition of lncRNA expression (false mRNA expression) to the suppressed mRNA expression may result in an opposite outcome that mRNAs seem to be expressed. Therefore, the lncRNA-error must be accounted for in mRNA-only quantification.

To reduce this lncRNA-error rate, we also investigated error rate in mRNA-lncRNA quantification pipeline. Using the mRNA and lncRNA annotation effectively reduced the lncRNA-error rate, which is almost at zero. However, it induced higher TER, which is caused by the higher mRNA-error rate. The higher mRNA-error rate obscures transcript-based quantification more than gene-based

quantification. The 5’ and 3’ exon expression levels are crucial for transcript-based quantification; yet, those counts are frequently lost due to the overlaps in those regions. Despite this, the effect mRNA-error is still minute compared to that of the lncRNA-error because transcript-level quantification is rarely employed. The limitations of transcript-level expression profiling and the interest in gene functions lead to a more frequent use of gene-level profiling. Moreover, the undetected 5’ or 3’ exon expressions are all same in the individual samples. As the gene expressions are compared between samples, the undetected expression rarely affects comparing the gene expression between samples. Thus, mRNA-errors have minor effect on detecting the gene-level differential expression between samples. In summary, the mRNA-lncRNA pipeline is more beneficial as its high mRNA-error is less crucial in mRNA expression profiling. With that said, using lncRNA annotation should not go overlooked, even if the study interest is only on the mRNA profiling.

The validity of lncRNA annotation is crucial for correct mRNA-lncRNA quantification. Despite its importance, the annotated lncRNAs in seven DBs showed a higher variability compared to that of mRNA. Several lncRNA characteristics, like the number of annotated lncRNA and the lengths of lncRNA exons, are DB-specific. As the lncRNA study is a more recent topic compared to the mRNA, such inconsistent and missing annotation between lncRNA DBs is reasonable, and therefore the DBs themselves are less reliable. To give an illustration, the DB size of the ELA and RLA are significantly smaller than that of the NONCODE suggesting there are missing annotations. Also, there may be falsely annotated lncRNAs in some databases; some of the lncRNAs have exactly matching sequences with that of the mRNA (Supplementary Table 4.3). With this in mind, the DB inconsistency of lncRNA is expected to affect the overlap number, length, direction, and shape. The number of

annotated lncRNAs showed high correlation with the number of overlap region. In addition, overlap direction and shape had higher dissimilarity between lncRNA DBs than between mRNA DBs. All in all, while the unpolished nature of lncRNA annotation have influence in the quantification step, proper lncRNA annotation will provide support for correct quantification of RNAs.

To reduce error rate, several error-reducing methods that are reasonably effective were examined. In the sequencing steps, the use of the strand-specific library is crucial for minimizing the TER, which is based on our observation that different-strand overlaps are more abundant than those in the same strand. The strand specific library along with specific aligners and quantifiers is recommended for minimized errors. In the analysis steps, some tools for alignment and quantification showed higher performance in terms of error rate. Hisat2-featureCounts combination produced the lowest TER in gene-based quantification, and such genome reference-based tools are recommended. Under transcript-based quantification, we suggest using eXpress over the most concurrent tool, Salmon. While the Salmon and eXpress had their pros and cons, Salmon has multi-mapping issues that can even increase the errors and inadequate expression measures. In summary, the more accurate measurement of expression is possible by using such methods.

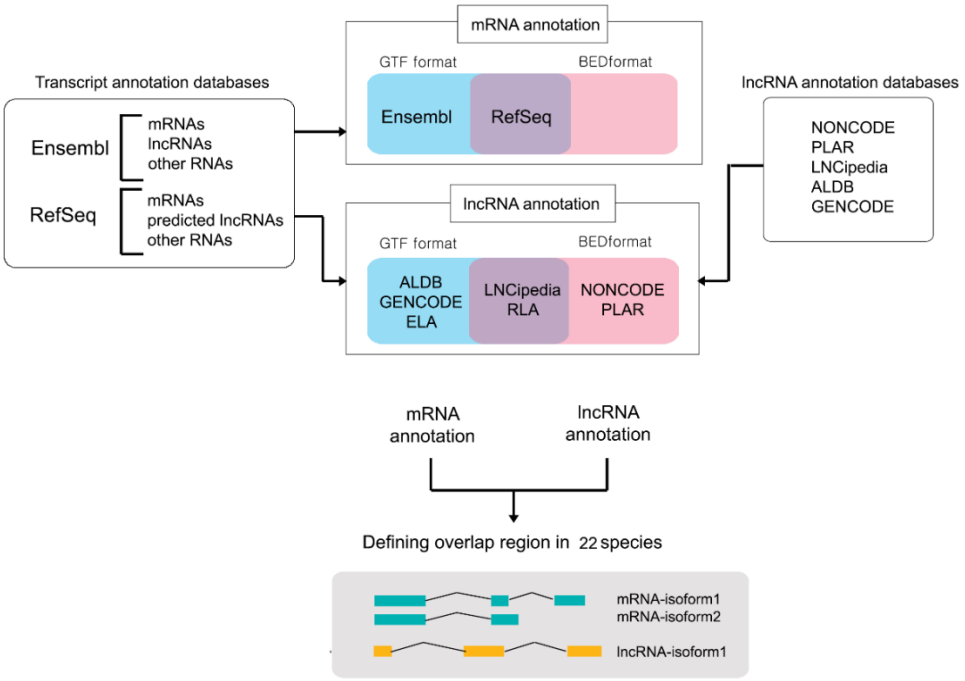
Although many tools are used for error-reducing purposes, these tools are not optimized to quantify overlap region with long read or fragment. In second generation sequencing, the long read and fragment amplify the errors with the current analysis tools. As for the third generation sequencing, the TER significantly decreases, with some drawback on lncRNA-error rate in mRNA-lncRNA quantification. The stringent quantification algorithm for the reads in the overlap regions might be the underlying reason. Without this optimization, long read or fragment could otherwise

amplify the error. This suggests a need for a quantification tool, which can effectively reduce the error caused by the overlap region, longer read and fragment. By regulating featureCounts overlap option, we tried to solve the defective quantification of longer reads. A less stringent quantification of longer reads from the overlap region seems to reduce and effectively control the error. An algorithm that can consider the read length and have flexibility in the overlap region quantification is desired for future RNA expression profiling.

As overlap error is species specific, overlap error needs to be considered according to different species. The overlap error is species-specific; this is directly related to the fact that overlap characteristics are species-specific. One of the characteristics, which affects error rate, is the percentage of overlapped transcript. The percentage of overlapped transcripts could be represented by the number of overlapped pairs and their transcript density to consider the different genome length between species. Non-primates—A.Thaliana, C.elegans, and Fruitfly—show a low number of overlapped pairs compared to their transcript density. Considering that the overlapped mRNA and lncRNA could regulate transcription, we can estimate that non-primates are less affected by this regulation than primates. Other characteristics such as overlap shape can cause species-specific error. In the simulation of error related to number of overlap and overlap length, Fruitfly showed a high variance of error compared to Human and Mouse. In summary, different overlap characteristics in various species could affect the degree of error. Summarizing these overlap characteristics in 22 species may help adequately quantifying the transcriptome in various species.

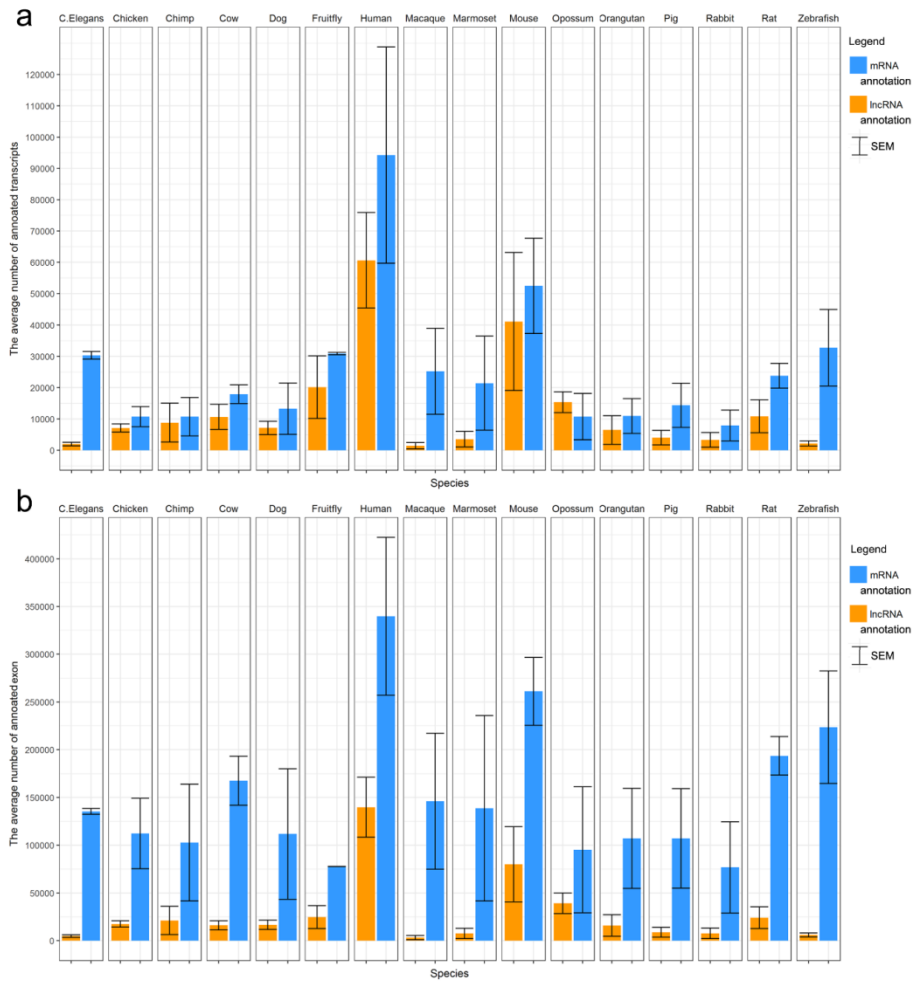
All in all, we considered several angles of the errors caused by mRNA-lncRNA overlap in current RNA quantification and displayed the importance of error

reduction in transcriptome analysis. Since these errors have a crucial role in measuring the abundance of transcripts, they could influence normalization and statistical analysis for differentially expressed genes (DEG). As DEGs are used to explain the change in traits according to the gene expression profiles, exact quantification of gene expression is important. A higher TER could influence the gene expression comparison; as the TER includes a high ratio of unquantified reads, a normalization method that uses total read counts could be affected. Also, an increased number of quantified transcripts, by lncRNA annotation, could make the significance level threshold more stringent for multiple testing correction. Thus, using reliable annotation and improving quantification algorithm is needed to improve the quantification in transcriptome study.



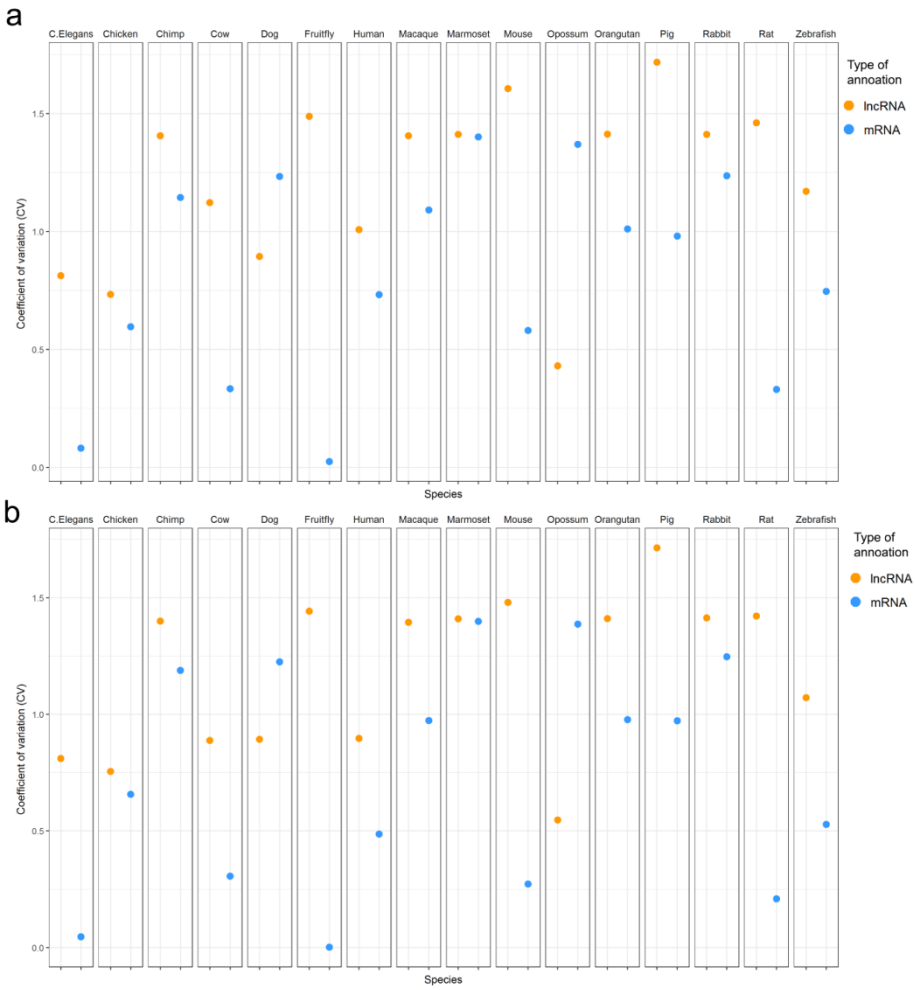
Supplementary Figure 4.1.Workflow of defining the overlap region in various databases.

The mRNA and lncRNA annotation from all available database were downloaded and analysed for overlap region.



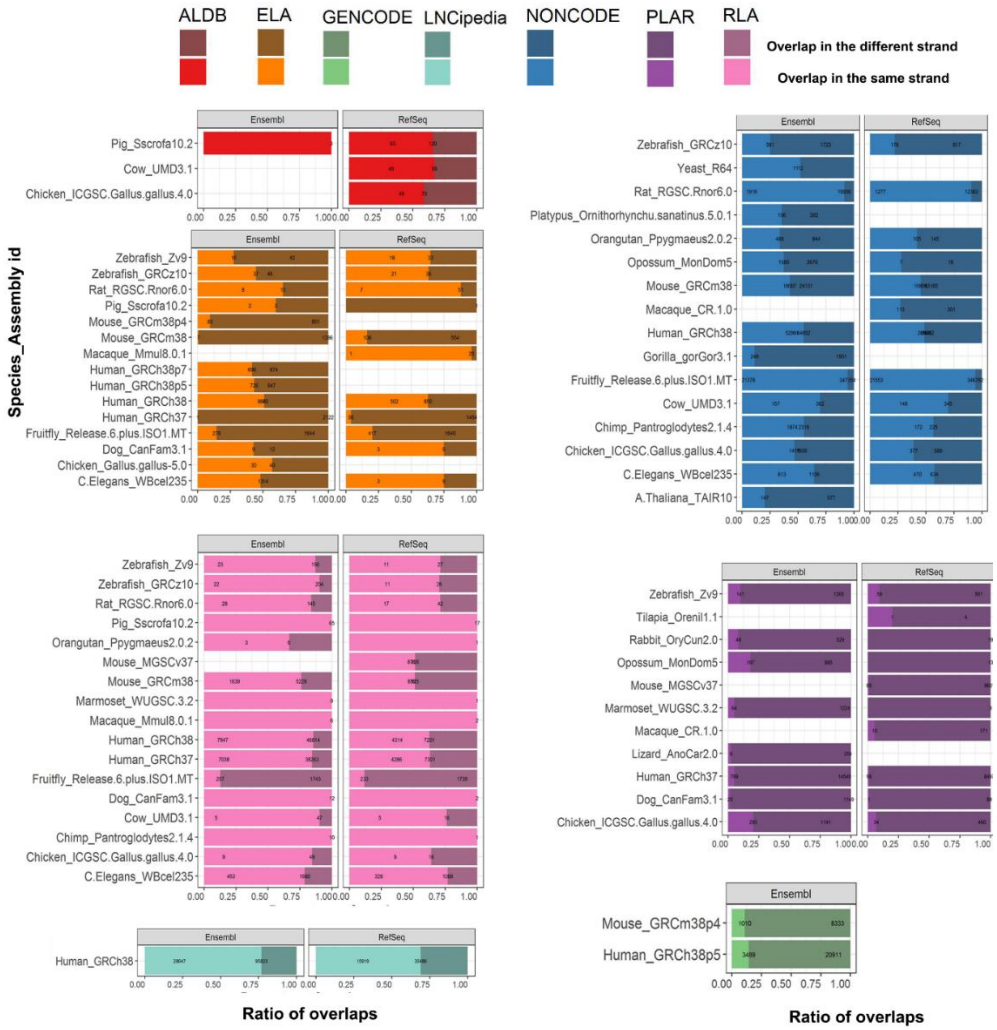
Supplementary Figure 4.2.The average number of annotation features.

The y-axis represents the average of annotation features. Among 23 species, 16 species were selected to measure variability among the annotation databases. In case of specific species that have several versions of reference genomes, the most representative genomes were selected based on the highest number of annotation databases. Accordingly, in Human and Zebrafish, GRCh38 and GRCz10 were respectively used, because they have more database source compared with other genome references. A) The average number of transcripts. Human has the most number of transcripts in both mRNA and lncRNAs. B) The average number of exons. The number of exons showed larger differences between the average of mRNA and lncRNA.



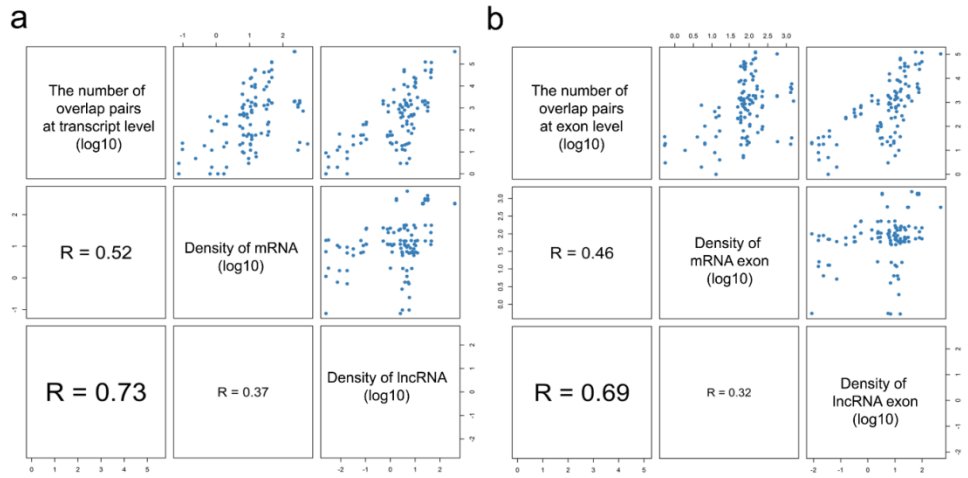
Supplementary Figure 4.3. Coefficient of variation in annotation counts between databases.

The coefficient of variation (CV) was calculated by dividing the standard deviation value between databases by the mean. a) CV value at the transcript level. When CV values were compared between types of RNA, they were higher in mRNA databases compared to that of lncRNA databases at all levels. Except for dog and opossum, the tendency of higher CV in lncRNA databases is displayed in most species. b) CV value at exon level. The value of CV at transcript levels are similar to those of exon level.



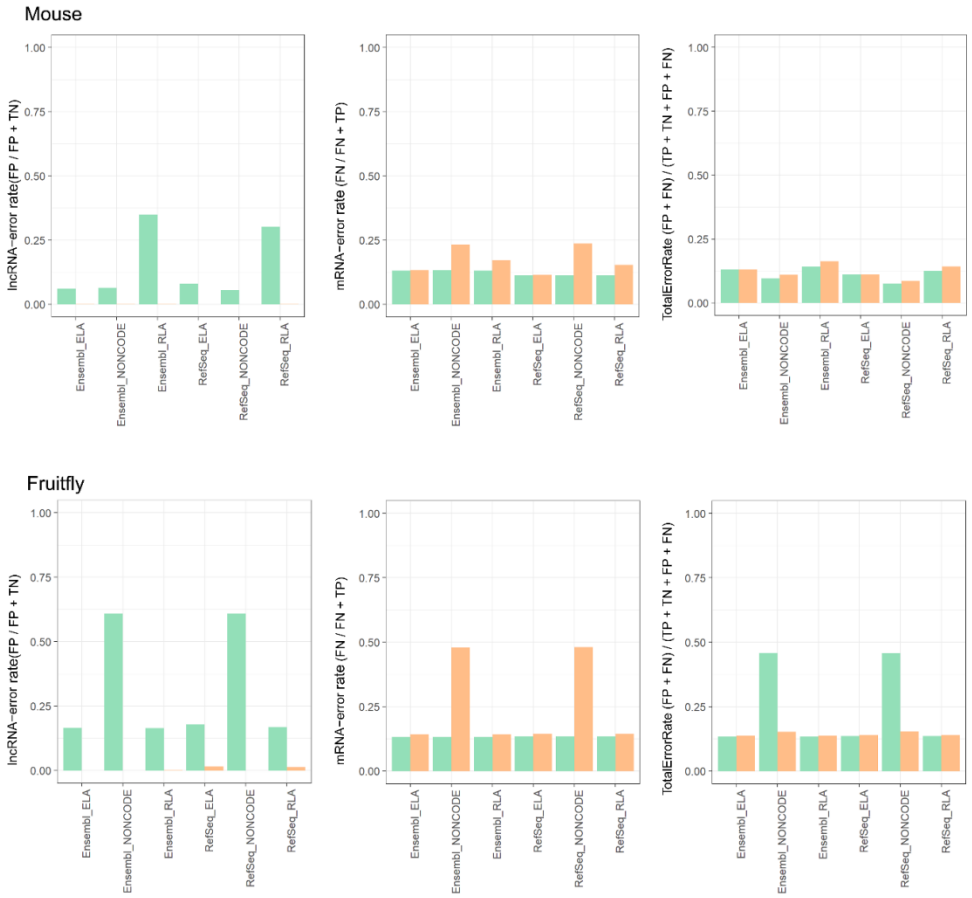
Supplementary Figure 4.4. Ratio of overlaps in different and same strand

The mRNA-lncRNA overlap can be found in same or different strand; 1) overlap in same strand and 2) overlap in different strand. The percentage of overlap is defined as the number of overlap pair in same (or different) strand over the total number of overlapped pairs.



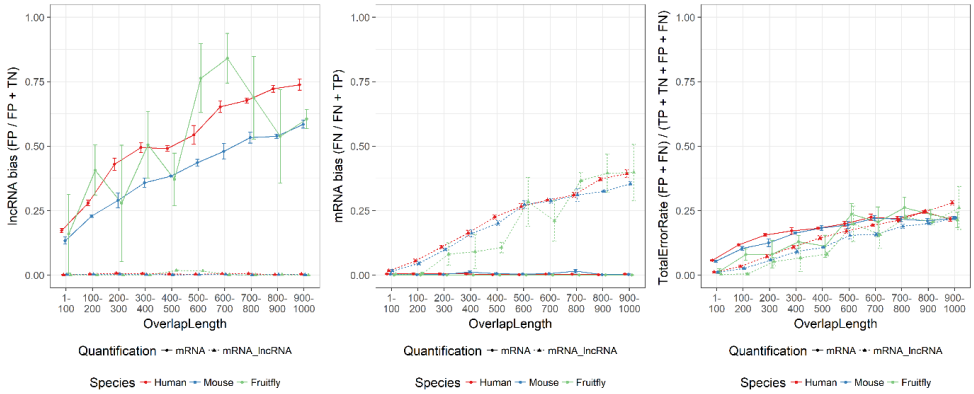
Supplementary Figure 4.5. Correlation plot of annotation density and the number of overlapped pairs.

The annotation trait (density of transcripts) and overlap trait (number) were considered. A) In the case of transcripts, the transcript density based on the number of transcripts over total genome length was used. Here, transcript density represents the number of transcripts per 1,000,000 bp of the genome. All trait combinations display a positive correlation according to the plot. B) In the case of exons, the exon density based on the number of exons over total genome length was used. Here, exon density represents the number of exons per 1,000,000 bp of the genome. All trait combinations display a positive correlation according to the plot. Among all, the number of overlap at transcript showed the highest correlation with the density of lncRNA annotation.



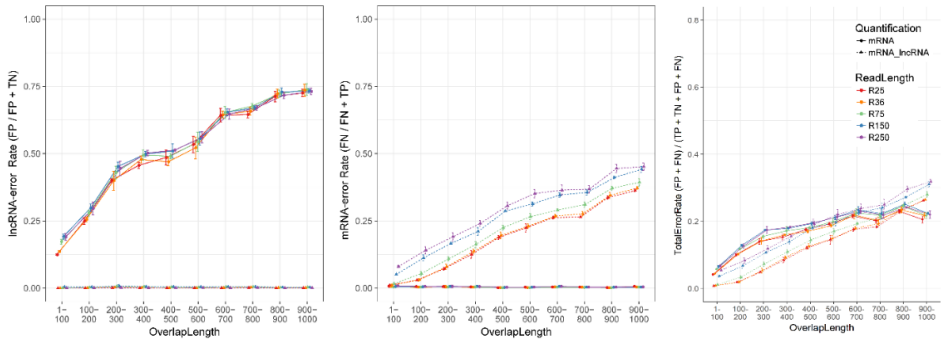
Supplementary Figure 4.6. Error plots of whole transcript simulation in combinations of Mouse and Fruitfly databases.

The barplot illustrates the mRNA error, lncRNA error, and total error rate in all database combinations for Mouse and Fruitfly. The mRNA-only and mRNA-lncRNA quantification are respectively colored in green and orange.



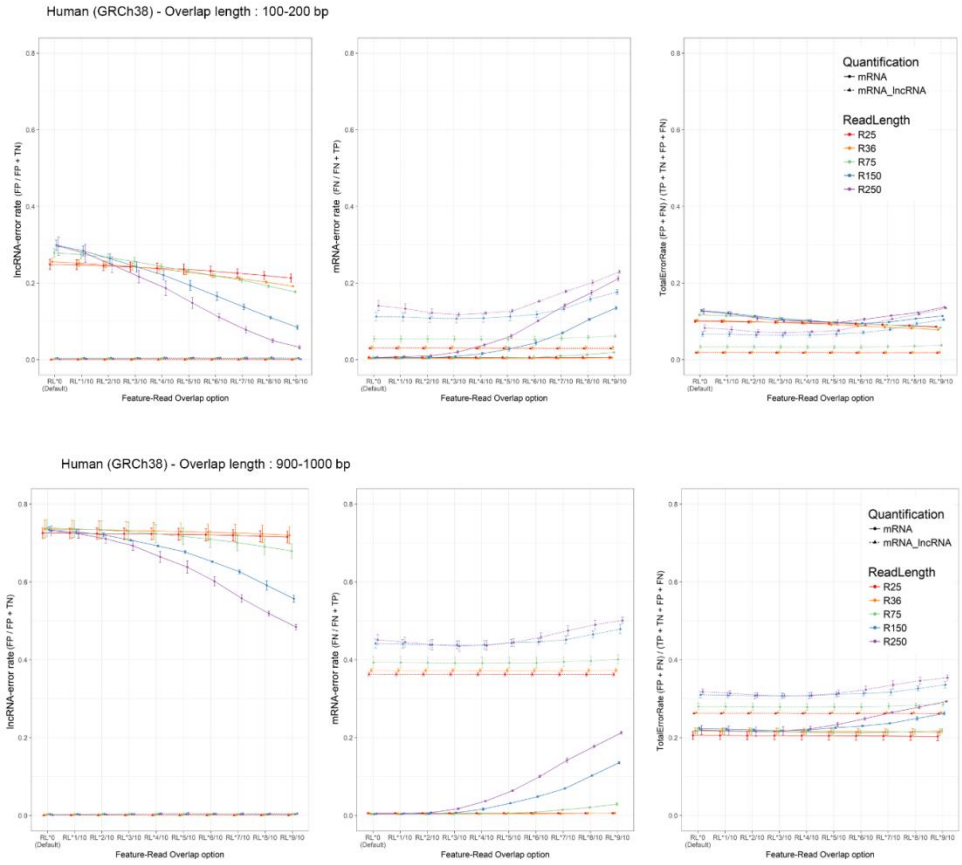
Supplementary Figure 4.7. Plot of error simulation according to overlap length.

The total Error rate is used here. Human, Mouse, and Fruitfly, are color-coded (red, blue, and green). The solid and dashed lines correspond to the mRNA-only and lncRNA-mRNA quantification, respectively.

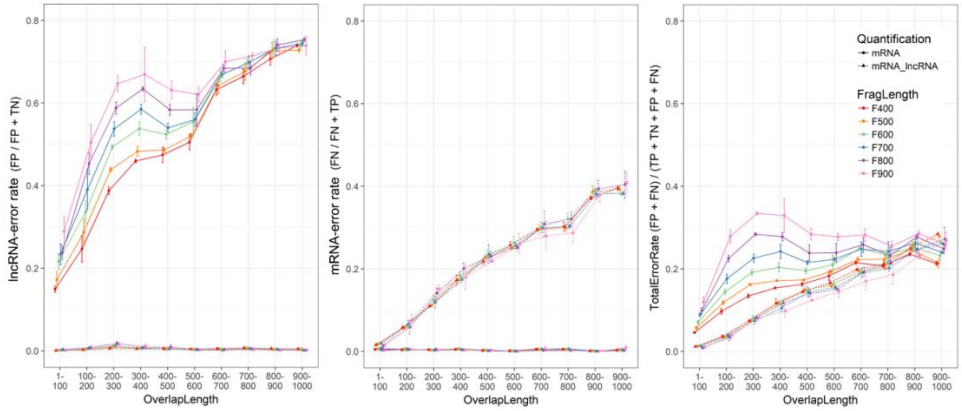


Supplementary Figure 4.8. Plot of error simulation according to read length in Human.

TER was presented. The relationship between read length and overlap length in the error rate. The different read length from 25 to 250 bp were color-coded.

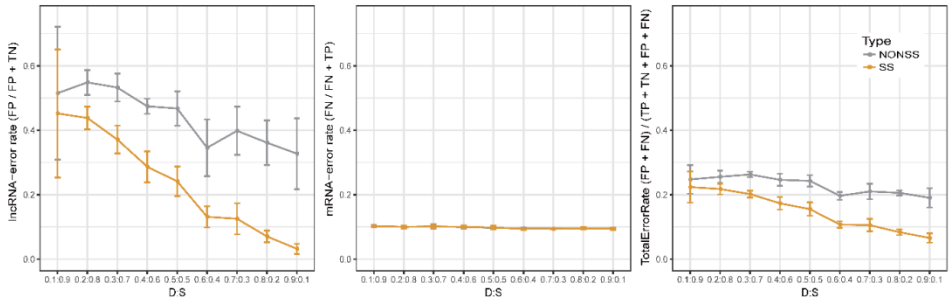


Supplementary Figure 4.9. Plot of error according to read length and feature-read overlap option



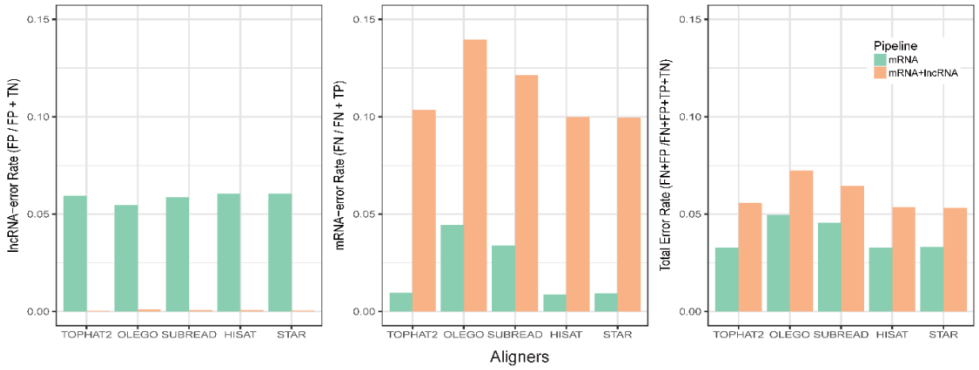
Supplementary Figure 4.10. Simulation for error according to fragment length in Human.

TER was presented. The relationship between read length and overlap length in the error rate. The different read length from 25 to 250 bp were color-coded.



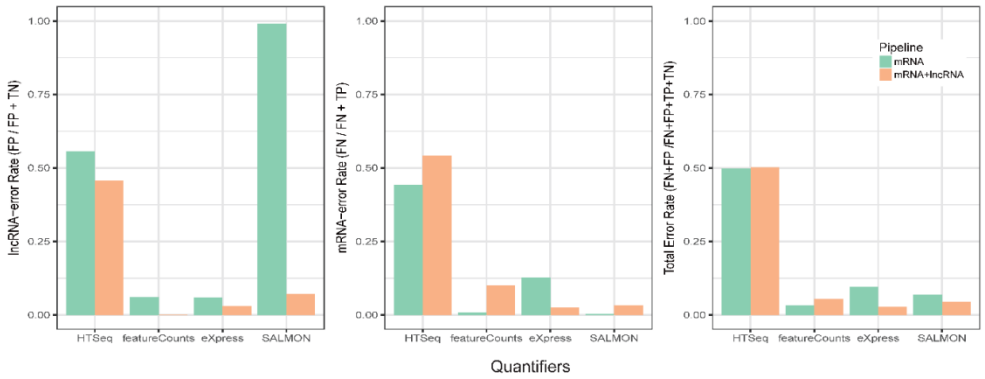
Supplementary Figure 4.11. The error comparison between strand-specific library vs. non-strand-specific library under different ratio of overlap direction.

The error rates in mRNA quantification were illustrated. D: proportion of different-strand overlap and S: proportion of same-strand overlap. The sum of each D:S ratio is 1.



Supplementary Figure 4.12. The error rates between five different aligners (Tophat2, Olego, Subread, Hisat2, and STAR).

Three error rates are plotted; lncRNA-error, mRNA-error, and total error. The barplots have been color-coded by the quantification pipelines: mRNA in green and mRNA+lncRNA in orange.



Supplementary Figure 4.13. The error rates between four quantifiers, two aligner-independent (HTSeq and featureCounts) and two aligner-dependent (eXpress and Salmon) quantifiers.

Three error rates are plotted; lncRNA-error, mRNA-error, and total error. The barplots have been color-coded by the quantification pipelines: mRNA in green and mRNA+lncRNA in orange.

Supplementary Table 4.1. Experimental design for lncRNA annotation and matched mRNA annotation

ELA and RLA are Ensembl and RefSeq lncRNA annotation, respectively. “O” and “X” respectively represents the existence and absence of DB for the corresponding species. Common names for species in Ensembl were used. Here, the species were listed in the order of the number of Assembly. The number of annotation data in each database indicates the number of “O” based on each annotation database. The number of combinations were calculated by multiplying the number of lncRNA annotations by that of lncRNA in the same reference assembly id.

No.	Species	Assembly ID	mRNA Annotation		lncRNA Annotation							# of combination
			RefSeq	Ensembl	RLA	ELA	NONCODE	PLAR	ALDB	GENCODE	LNCipedia	
1	Human	GRCh37	O	O	O	O	X	O	X	X	X	6
		GRCh38	O	O (p5)	O	O (p5)	O	X	X	O (p5)	O	10
2	Mouse	MGSCv37	O	X	O	X	X	O	X	X	X	2
		GRCm38	O	O (p4)	O	O (p4)	O	X	X	O (p4)	X	8
3	Chicken	ICGSC Gallus_gallus-4.0	O	O	O	X	O	O	O	X	X	8
		Gallus_gallus-5.0	X	O	X	O	X	X	X	X	X	1
4	Macaque	CR_1.0	O	X	O	X	O	O	X	X	X	3
		Mmul_8.0.1	O	O	O	O	X	X	X	X	X	4

5	Zebrafish	Zv9	O	O	O	O	X	O	X	X	X	6
		GRCz10	O	O	O	O	O	X	X	X	X	6
6	A.Thaliana	TAIR10	X	O	X	X	O	X	X	X	X	1
7	C.Elegans	WBcel235	O	O	O	O	O	X	X	X	X	6
8	Chimp	Pan_troglodytes-2.1.4	O	O	O	X	O	X	X	X	X	4
9	Cow	UMD_3.1	O	O	O	X	O	X	O	X	X	6
10	Dog	CanFam3.1	O	O	O	O	X	O	X	X	X	6
11	Fruitfly	Release 6 plus ISO1 MT	O	O	O	O	O	X	X	X	X	6
12	Gorilla	gorGor3.1	X	O	X	X	O	X	X	X	X	1
13	Lizard	AnoCar2.0	X	O	X	O	X	O	X	X	X	2
14	Marmoset	WUGSC 3.2	O	O	O	X	X	O	X	X	X	4
15	Opossum	MonDom5	O	O	X	X	O	O	X	X	X	4
16	Orangutan	P_pygmaeus_2.0.2	O	O	O	X	O	X	X	X	X	4
17	Pig	Sscrofa10.2	O	O	O	O	X	X	O	X	X	6
18	Platypus	Ornithorhynchus_anatinus-5.0.1	O	O	X	X	O	X	X	X	X	2
19	Rabbit	OryCun2.0	O	O	O	X	X	O	X	X	X	4
20	Rat	RGSC Rnor_6.0	O	O	O	O	O	X	X	X	X	6
21	Tilapia	Orenil1.1	O	X	X	X	X	O	X	X	X	1
22	Yeast	R64	X	O	X	X	O	X	X	X	X	1
Sum			22	27	19	16	16	11	3	2	1	119

References

Aken, B. L., et al. (2016). "Ensembl 2017." Nucleic acids research **45**(D1): D635-D642.

Altmann, A., et al. (2007). "Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance." Antiviral therapy **12**(2): 169.

Anders, S., et al. (2015). "HTSeq--a Python framework to work with high-throughput sequencing data." Bioinformatics **31**(2): 166-169.

Ayers, K. L., et al. (2015). "Identification of candidate gonadal sex differentiation genes in the chicken embryo using RNA-seq." BMC genomics **16**(1): 704.

Baxter, N. T., et al. (2016). "DNA from fecal immunochemical test can replace stool for detection of colonic lesions using a microbiota-based model." Microbiome **4**(1): 59.

Baxter, N. T., et al. (2016). "Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions." Genome Med **8**(1): 37.

Belkaid, Y. and T. W. Hand (2014). "Role of the microbiota in immunity and inflammation." Cell **157**(1): 121-141.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the royal statistical society. Series B (Methodological): 289-300.

Bianchini, F. and H. Vainio (2001). "Allium vegetables and organosulfur compounds: do they help prevent cancer?" Environmental health perspectives **109**(9): 893.

Biddle, A., et al. (2013). "Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities." Diversity **5**(3): 627-640.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.

Bolger, A. M., et al. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics: btu170.

Bolstad, B. M., et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.

Borodina, T., et al. (2011). A strand-specific library preparation protocol for RNA sequencing. Methods in enzymology, Elsevier. **500:** 79-98.

Bu, D., et al. (2011). "NONCODE v3. 0: integrative annotation of long noncoding RNAs." Nucleic acids research: gkr1175.

Cai, Y.-D., et al. (2006). "Using LogitBoost classifier to predict protein structural classes." Journal of theoretical biology **238**(1): 172-176.

Callahan, B. J., et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data." Nat Methods **13**(7): 581-583.

Caporaso, J. G., et al. (2010). "QIIME allows analysis of high-throughput community sequencing data." Nature methods **7**(5): 335.

Carbonero, F., et al. (2012). "Microbial pathways in colonic sulfur metabolism and links with health and disease." Front Physiol **3**: 448.

Cervantes, H. M. (2015). "Antibiotic-free poultry production: is it sustainable?" Journal of Applied Poultry Research **24**(1): 91-97.

Chang, C.-D., et al. (2011). "Using data mining techniques for multi-diseases prediction modeling of hypertension and hyperlipidemia by common risk factors." Expert Systems with Applications **38**(5): 5507-5513.

Chen, L. L. and G. G. Carmichael (2010). "Decoding the function of nuclear long non-coding RNAs." Curr Opin Cell Biol **22**(3): 357-364.

Cho, S.-B. and H.-H. Won (2003). Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19, Australian Computer Society, Inc.

Conlon, M. A. and A. R. Bird (2014). "The impact of diet and lifestyle on gut microbiota and human health." Nutrients **7**(1): 17-44.

Cruz, J. A. and D. S. Wishart (2006). "Applications of machine learning in cancer prediction and prognosis." Cancer informatics **2**.

Cutler, R. and P. Wilson (2004). "Antibacterial activity of a new, stable, aqueous extract of allicin against methicillin-resistant Staphylococcus aureus." British journal of biomedical science **61**(2): 71-74.

D'Amelio, P. and F. Sassi (2018). "Gut Microbiota, Immune System, and Bone." Calcif Tissue Int **102**(4): 415-425.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. ICML, Citeseer.

David, L. A., et al. (2014). "Diet rapidly and reproducibly alters the human gut microbiome." nature **505**(7484): 559-563.

Derrien, T., et al. (2012). "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression." Genome Res **22**(9): 1775-1789.

DeSantis, T. Z., et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Applied and environmental microbiology **72**(7): 5069-5072.

Di Paola, M., et al. (2016). "Alteration of Fecal Microbiota Profiles in Juvenile Idiopathic Arthritis. Associations with HLA-B27 Allele and Disease Status." Front Microbiol **7**: 1703.

Dinh, D. M., et al. (2014). "Intestinal microbiota, microbial translocation, and systemic inflammation in chronic HIV infection." The Journal of infectious diseases **211**(1): 19-27.

Eddy, S. R. (2001). "Non–coding RNA genes and the modern RNA world." Nature Reviews Genetics **2**(12): 919-929.

Edwards, C., et al. (2017). "Polyphenols and health: Interactions between fibre, plant polyphenols and the gut microbiota." Nutrition bulletin **42**(4): 356-360.

Eeckhaut, V., et al. (2012). "Butyricicoccus pullicaecorum in inflammatory bowel disease." Gut: gutjnl-2012-303611.

Esteller, M. (2011). "Non-coding RNAs in human disease." Nat Rev Genet **12**(12): 861-874.

Eun Byeol Lee, S.-H. L., Seung-Hwan Kim, Sang-Hyun Kang, Kyung-Woo Lee, Da-Hye Kim, Dong-Wook Kim, Hwan-Gu Kang, Nam-Seok Kim, Jung-Bong Kim, Jung-Suk Choe, Hwan-Hee Jang, You-Jin Hwang, You-Suk Kim, Sung-Hyen Lee "Effects of Dietary Allium Hookeri on Growth and Blood Biochemical Parameters in Broiler Chickens." Kor. J. Pharmacogn **In press**.

Ferrario, C., et al. (2017). "How to Feed the Mammalian Gut Microbiota: Bacterial and Metabolic Modulation by Dietary Fibers." Front Microbiol **8**: 1749.

Ferrario, C., et al. (2014). "Modulation of Fecal Clostridiales Bacteria and Butyrate by Probiotic Intervention with Lactobacillus paracasei DG Varies among Healthy Adults–3." The Journal of nutrition **144**(11): 1787-1796.

Franco-Zorrilla, J. M., et al. (2007). "Target mimicry provides a new mechanism for regulation of microRNA activity." Nat Genet **39**(8): 1033-1037.

Friedman, J., et al. (2000). "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." The annals of statistics **28**(2): 337-407.

Fujisawa, H., et al. (2009). "Antibacterial potential of garlic-derived allicin and its cancellation by sulfhydryl compounds." Bioscience, biotechnology, and biochemistry **73**(9): 1948-1955.

Garalde, D. R., et al. (2018). "Highly parallel direct RNA sequencing on an array of nanopores." Nature methods **15**(3): 201.

Gerritsen, J., et al. (2011). "Intestinal microbiota in human health and disease: the impact of probiotics." Genes & nutrition **6**(3): 209.

Giloteaux, L., et al. (2016). "Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome." Microbiome **4**(1): 30.

Gong, Y., et al. (2017). "lncRNA-screen: an interactive platform for computationally screening long non-coding RNAs in large genomics datasets." BMC genomics **18**(1): 434.

Goodrich, J. K., et al. (2014). "Human genetics shape the gut microbiome." Cell **159**(4): 789-799.

Griebel, T., et al. (2012). "Modelling and simulating generic RNA-Seq experiments with the flux simulator." Nucleic acids research **40**(20): 10073-10083.

Guil, S. and M. Esteller (2012). "Cis-acting noncoding RNAs: friends and foes." Nature structural & molecular biology **19**(11): 1068-1075.

Guttman, M., et al. (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**(7235): 223-227.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." The Journal of Machine Learning Research **3**: 1157-1182.

Heo, J., et al. (2016). "Gut microbiota Modulated by Probiotics and Garcinia cambogia Extract Correlate with Weight Gain and Adipocyte Sizes in High Fat-Fed Mice." Scientific reports **6**.

Heo, J. B., et al. (2013). "Epigenetic regulation by long noncoding RNAs in plants." Chromosome Res **21**(6-7): 685-693.

Hezroni, H., et al. (2015). "Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species." Cell Rep **11**(7): 1110-1122.

Hornik, K., et al. (2007). "RWeka: an R interface to Weka." R package version 0.3-2.

Hsu, C.-W. and C.-J. Lin (2002). "A comparison of methods for multiclass support vector machines." IEEE transactions on Neural Networks **13**(2): 415-425.

Hubbard, T., et al. (2002). "The Ensembl genome database project." Nucleic acids research **30**(1): 38-41.

Hunter, S., et al. (2014). "EBI metagenomics—a new resource for the analysis and archiving of metagenomic data." Nucleic acids research **42**(D1): D600-D606.

Huson, D. H., et al. (2007). "MEGAN analysis of metagenomic data." Genome research **17**(3): 377-386.

Hwang, J.-S., et al. (2015). "Total Phenolics, Total Flavonoids, and Antioxidant Capacity in the Leaves, Bulbs, and Roots of Allium hookeri." Korean Journal of Food Science and Technology **47**(2): 261-266.

Jain, A. and D. Zongker (1997). "Feature selection: Evaluation, application, and small sample performance." Pattern Analysis and Machine Intelligence, IEEE Transactions on **19**(2): 153-158.

Jangi, S., et al. (2016). "Alterations of the human gut microbiome in multiple sclerosis." Nat Commun **7**: 12015.

Jiao, Y., et al. (2010). "Predictive models of autism spectrum disorder based on brain regional cortical thickness." Neuroimage **50**(2): 589-599.

Kapranov, P., et al. (2007). "RNA maps reveal new RNA classes and a possible function for pervasive transcription." Science **316**(5830): 1484-1488.

Kapranov, P., et al. (2007). "Genome-wide transcription and the implications for genomic organization." Nat Rev Genet **8**(6): 413-423.

Karolchik, D. (2003). "The UCSC Genome Browser Database." Nucleic acids research **31**(1): 51-54.

Kashyap, P. C., et al. (2017). Microbiome at the frontier of personalized medicine. Mayo Clinic Proceedings, Elsevier.

Katayama, S., et al. (2005). "Antisense transcription in the mammalian transcriptome." Science **309**(5740): 1564-1566.

Kawaji, H., et al. (2014). "Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing." Genome Res **24**(4): 708-717.

Kim, C.-H., et al. (2012). "Anti-inflammatory effect of Allium hookeri root methanol extract in LPS-induced RAW264. 7 cells." Journal of the Korean Society of Food Science and Nutrition **41**(11): 1645-1648.

Kim, D., et al. (2015). "HISAT: a fast spliced aligner with low memory requirements." Nature methods **12**(4): 357-360.

Kim, D. K., et al. (2013). "Dietary Curcuma longa enhances resistance against Eimeria maxima and Eimeria tenella infections in chickens." Poultry science **92**(10): 2635-2643.

Kim, J.-H., et al. (2017). "Effect of Allium hookeri and whey powder in diet of pigs on physicochemical characteristics and oxidative stability of pork." Italian Journal of Animal Science: 1-9.

Kim, J. E., et al. (2015). "Dietary Capsicum and Curcuma longa oleoresins increase intestinal microbiome and necrotic enteritis in three commercial broiler breeds." Research in veterinary science **102**: 150-158.

Kim, K., et al. (2015). "Application of LogitBoost Classifier for Traceability Using SNP Chip Data." PLoS One **10**(10): e0139685.

Kim, N. S., Choi, B. K., Lee, S. H., Jang, H. H,, Kim, J. B., Kim, H. R., Kim, D. K., Kim, Y. S., Yang, J. H., Kim, H. J. and Lee, S. H. (2015). "Effects of Allium hookeri on glucose metabolism in type II diabetic mice." Kor J Pharmacogn **46**: 78-83.

Kino, T., et al. (2010). "Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor." Sci Signal **3**(107): ra8.

Kinross, J. M., et al. (2011). "Gut microbiome-host interactions in health and disease." Genome medicine **3**(3): 1.

Knights, D., et al. (2011). "Supervised classification of human microbiota." FEMS Microbiol Rev **35**(2): 343-359.

Kukar, M., et al. (1999). "Analysing and improving the diagnosis of ischaemic heart disease with machine learning." Artificial intelligence in medicine **16**(1): 25-50.

Kung, J. T., et al. (2013). "Long noncoding RNAs: past, present, and future." Genetics **193**(3): 651-669.

Kuo, S. M. (2013). "The interplay between fiber and the intestinal microbiome in the inflammatory response." Adv Nutr **4**(1): 16-28.

Landwehr, N., et al. (2005). "Logistic model trees." Machine Learning **59**(1-2): 161-205.

Langille, M. G., et al. (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences." Nat Biotechnol **31**(9): 814-821.

Lee, H. C., et al. (2006). "Effect of tea phenolics and their aromatic fecal bacterial metabolites on intestinal microbiota." Research in microbiology **157**(9): 876-884.

Lee, J.-H., et al. (2011). "rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries." The Journal of Microbiology **49**(4): 689.

Lee, K.-W., et al. (2014). "Comparison of Effect of Water and Ethanolic Extract from Roots and Leaves of Allium hookeri." Journal of the Korean Society of Food Science and Nutrition **43**(12): 1808-1816.

Lee, S. H., Kim, N. S., Choi, B. K., Jang, H. H., Kim, J. B., Lee, Y. M., Kim, D. K., Lee, C. H., Kim, Y. S., Yang, J. H., Kim, Y. S., Kim, H. J. and Lee, S. H (2015). "Effects of Allium hookeri on lipid metabolism in type II diabetic mice." Kor. J. Pharmacogn **46**: 148-153.

LEE, Y.-K. (2013). "Effects of diet on gut microbiota profile and the implications for health and disease." Bioscience of microbiota, food and health **32**(1): 1-12.

Lee, Y., et al. (2017). "Dietary Allium hookeri reduces inflammatory response and increases expression of intestinal tight junction proteins in LPS-induced young broiler chicken." Research in veterinary science **112**: 149-155.

Lee, Y., et al. (2017). "In vitro analysis of the immunomodulating effects of Allium hookeri on lymphocytes, macrophages, and tumour cells." The Journal of Poultry Science **54**(2): 142-148.

Levin, J. Z., et al. (2010). "Comprehensive comparative analysis of strand-specific RNA sequencing methods." Nat Methods **7**(9): 709-715.

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a

Li, A., et al. (2015). "ALDB: a domestic-animal long noncoding RNA database." PLoS One **10**(4): e0124003.

Li, B., et al. (2010). "RNA-Seq gene expression estimation with read mapping uncertainty." Bioinformatics **26**(4): 493-500.

Li, L., et al. (2018). "Infectious bursal disease virus infection leads to changes in the gut associated-lymphoid tissue and the microbiota composition." PLoS One **13**(2): e0192066.

Liao, Y., et al. (2013). "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features." Bioinformatics: btt656.

Lillehoj, H. S., et al. (2011). Effects of dietary plant-derived phytonutrients on the genome-wide profiles and coccidiosis resistance in the broiler chickens. BMC proceedings, BioMed Central.

Liu, Z., et al. (2011). "Sparse distance-based learning for simultaneous multiclass classification and feature selection of metagenomic data." Bioinformatics **27**(23): 3242-3249.

Love, M. I., et al. (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology **15**(12): 550.

Lozupone, C. and R. Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities." Appl Environ Microbiol **71**(12): 8228-8235.

Lv, Q., et al. (2016). "D-repeat in the XIST gene is required for X chromosome inactivation." RNA Biol **13**(2): 172-176.

Makalowska, I., et al. (2005). "Overlapping genes in vertebrate genomes." Comput Biol Chem **29**(1): 1-12.

Manach, C., et al. (2004). "Polyphenols: food sources and bioavailability." The American journal of clinical nutrition **79**(5): 727-747.

Maranduba, C. M. d. C., et al. (2015). "Intestinal microbiota as modulators of the immune system and neuroimmune system: impact on the host health and homeostasis." Journal of immunology research **2015**.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet. journal **17**(1): pp. 10-12.

Mattick, J. S. and I. V. Makunin (2006). "Non-coding RNA." Hum Mol Genet **15 Spec No 1**: R17-29.

.

Mercer, T. R., et al. (2009). "Long non-coding RNAs: insights into functions." Nature Reviews Genetics **10**(3): 155-159.

Mortazavi, A., et al. (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Nagai, F., et al. (2009). "Parasutterella excrementihominis gen. nov., sp. nov., a member of the family Alcaligenaceae isolated from human faeces." International journal of systematic and evolutionary microbiology **59**(7): 1793-1797.

Nencini, C., et al. (2007). "Evaluation of antioxidative properties of Allium species growing wild in Italy." Phytotherapy Research **21**(9): 874-878.

Noguera-Julian, M., et al. (2016). "Gut Microbiota Linked to Sexual Preference and HIV Infection." EBioMedicine **5**: 135-146.

Ohlsson, C. and K. Sjögren (2015). "Effects of the gut microbiota on bone mass." Trends in Endocrinology & Metabolism **26**(2): 69-74.

Oksanen, J. and F. G. Blanchet "Package 'vegan'."

Ozdal, T., et al. (2016). "The Reciprocal Interactions between Polyphenols and Gut Microbiota and Effects on Bioaccessibility." Nutrients **8**(2): 78.

Ozsolak, F. and P. M. Milos (2011). "RNA sequencing: advances, challenges and opportunities." Nat Rev Genet **12**(2): 87-98.

Pacifici, R. (2010). "The immune system and bone." Archives of biochemistry and biophysics **503**(1): 41-53.

Pasolli, E., et al. (2016). "Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights." PLoS Comput Biol **12**(7): e1004977.

Pelechano, V. and L. M. Steinmetz (2013). "Gene regulation by antisense transcription." Nat Rev Genet **14**(12): 880-893.

Poretsky, R., et al. (2014). "Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics." PLoS One **9**(4): e93827.

Pruitt, K. D., et al. (2012). "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy." Nucleic Acids Res **40**(Database issue): D130-135.

Ravussin, Y., et al. (2012). "Responses of gut microbiota to diet composition and weight loss in lean and obese mice." Obesity (Silver Spring) **20**(4): 738-747.

Reiter, J., et al. (2017). "Diallylthiosulfinate (Allicin), a Volatile Antimicrobial from Garlic (Allium sativum), Kills Human Lung Pathogenic Bacteria, Including MDR Strains, as a Vapor." Molecules **22**(10): 1711.

Roberfroid, M., et al. (2010). "Prebiotic effects: metabolic and health benefits." British Journal of Nutrition **104**(S2): S1-S63.

Roberts, A. and L. Pachter (2013). "Streaming fragment assignment for real-time analysis of sequencing experiments." Nat Methods **10**(1): 71-73.

Robinson, M. D., et al. (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics **26**(1): 139-140.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome biology **11**(3): R25.

Roh, S. S., et al. (2016). "Allium hookeri root protects oxidative stress-induced inflammatory responses and beta-cell damage in pancreas of streptozotocin-induced diabetic rats." BMC Complement Altern Med **16**: 63.

tion of HO-1 in the pancreas of STZ-induced diabetic rats.

Rooks, M. G. and W. S. Garrett (2016). "Gut microbiota, metabolites and host immunity." Nature Reviews Immunology **16**(6): 341.

Rooks, M. G. and W. S. Garrett (2016). "Gut microbiota, metabolites and host immunity." Nat Rev Immunol **16**(6): 341-352.

Sainani, G., et al. (1979). "Onion, garlic, and experimental atherosclerosis." Japanese heart journal **20**(3): 351-357.

Sajda, P. (2006). "Machine learning for detection and diagnosis of disease." Annu. Rev. Biomed. Eng. **8**: 537-565.

Sakamoto, M. and Y. Benno (2006). "Reclassification of Bacteroides distasonis, Bacteroides goldsteinii and Bacteroides merdae as Parabacteroides distasonis gen. nov., comb. nov., Parabacteroides goldsteinii comb. nov. and Parabacteroides merdae comb. nov." International journal of systematic and evolutionary microbiology **56**(7): 1599-1605.

Sanna, C. R., et al. (2008). "Overlapping genes in the human and mouse genomes." BMC genomics **9**(1): 1.

Santoso, U., et al. (1995). "Effect of dried Bacillus subtilis culture on growth, body composition and hepatic lipogenic enzyme activity in female broiler chicks." British Journal of Nutrition **74**(4): 523-529.

Scalbert, A., et al. (2002). "Absorption and metabolism of polyphenols in the gut and impact on health." Biomedicine & Pharmacotherapy **56**(6): 276-282.

Schena, M., et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467.

Schloss, P. D., et al. (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." Appl. Environ. Microbiol. **75**(23): 7537-7541.

Schuierer, S. and G. Roma (2016). "The exon quantification pipeline (EQP): a comprehensive approach to the quantification of gene, exon and junction expression from RNA-seq data." Nucleic Acids Res **44**(16): e132.

Sengupta, A., et al. (2004). "Allium vegetables in cancer prevention: an overview." Asian Pacific Journal of Cancer Prevention **5**(3): 237-245.

Sigurgeirsson, B., et al. (2014). "Analysis of stranded information using an automated procedure for strand specific RNA sequencing." BMC genomics **15**(1): 1.

Singh, R. K., et al. (2017). "Influence of diet on the gut microbiome and implications for human health." J Transl Med **15**(1): 73.

Sjögren, K., et al. (2012). "The gut microbiota regulates bone mass in mice." Journal of Bone and Mineral Research **27**(6): 1357-1367.

Song, E.-Y., et al. (2014). "Effect of addition of Allium hookeri on the quality of fermented sausage with meat from sulfur fed pigs during ripening." Korean journal for food science of animal resources **34**(3): 263.

Statnikov, A., et al. (2005). "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis." Bioinformatics **21**(5): 631-643.

Sun, H., et al. (2015). "IAOseq: inferring abundance of overlapping genes using RNA-seq data." BMC bioinformatics **16**(Suppl 1): S3.

Tan, Z., et al. (2018). "Differences in gut microbiota composition in finishing Landrace pigs with low and high feed conversion ratios." Antonie Van Leeuwenhoek.

Tyner, C., et al. (2016). "The UCSC genome browser database: 2017 update." Nucleic acids research **45**(D1): D626-D634.

ur Rahman, S., et al. (2017). "In vivo effects of Allium cepa L. on the selected gut microflora and intestinal histomorphology in broiler." Acta histochemica **119**(5): 446-450.

Viveros, A., et al. (2011). "Effects of dietary polyphenol-rich grape products on intestinal microflora and gut morphology in broiler chicks." Poultry science **90**(3): 566-578.

Volders, P. J., et al. (2015). "An update on LNCipedia: a database for annotated human lncRNA sequences." Nucleic Acids Res **43**(Database issue): D174-180.

Walker, A. W., et al. (2011). "Dominant and diet-responsive groups of bacteria within the human colonic microbiota." ISME J **5**(2): 220-230.

Wang, K. C. and H. Y. Chang (2011). "Molecular mechanisms of long noncoding RNAs." Mol Cell **43**(6): 904-914.

Wang, Z., et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nature Reviews Genetics **10**(1): 57-63.

White, J. R., et al. (2009). "Statistical methods for detecting differentially abundant features in clinical metagenomic samples." PLoS computational biology **5**(4): e1000352.

Wu, Z. and M. J. Aryee (2010). "Subset quantile normalization using negative control features." Journal of Computational Biology **17**(10): 1385-1395.

Yang, Y., et al. (2015). "Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein." Cell Biosci **5**: 59.

Yates, A., et al. (2016). "Ensembl 2016." Nucleic Acids Res **44**(D1): D710-716. (http://github.com/Ensembl) under an Apache 2.0 license.

Yelin, R., et al. (2003). "Widespread occurrence of antisense transcription in the human genome." Nat Biotechnol **21**(4): 379-386.

Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. ICML.

Zeller, G., et al. (2014). "Potential of fecal microbiota for early-stage detection of colorectal cancer." Mol Syst Biol **10**: 766.

Zeng, B., et al. (2015). "The bacterial communities associated with fecal types and body weight of rex rabbits." Sci Rep **5**: 9342.

Zhang, J., et al. (2014). "PEAR: a fast and accurate Illumina Paired-End reAd mergeR." Bioinformatics **30**(5): 614-620.

Zhao, Y., et al. (2016). "NONCODE 2016: an informative and valuable data source of long non-coding RNAs." Nucleic Acids Res **44**(D1): D203-208.

요약(국문초록)

전사체와 메타지놈 데이터 분석을 위한 통계적 방법의 활용

방소현

협동과정 생물정보학 전공

서울대학교 대학원

시퀀싱 기술의 발달로 유전체, 전사체, 단백체, 후성 유전체, 메타지노믹와 같은 분야에서 유전체 단위로 생명체의 정보를 해독할 수 있게 되었다. 이 중 전사체와 메타지놈 분야는 정량화된 유전 정보를 다룬다는 공통점 때문에 통계적 분석 방법들을 공유하고 있다. 전사체 분석은 유전자들의 발현을 정량화 하고, 특정 조건 하에 다른 양으로 발현되는 유전자를 발굴하는 것을 목표하고 있다. 정량화된 양을 그룹 간 비교하는 거나 형질과 관련된 유전자를 찾기 위해 여러 통계적 분석 도구 및 방법들이 개발되었으며 널리 사용되고 있다. 메타지놈 분석은 정량화 하는 대상이 미생물들의 양이라는 것은 다르나 정량화한 미생물들의 양을 분석하기 때문에 전사체에서 사용되었던 방법들이 대부분 이용되고 있다.

양적 자료의 기본적인 분석에서 더 나아가서 머신러닝기법을 이용하여 정량화된 유전물질의 양으로 질병과 같은 형질을 예측하고자 하는 시도도 이루어 지고 있다. 특히, 인간의 장내미생물은 면역체계와 연관성이 있기 때문에 장내미생물의 종류와 양으로 질병을 진단하려는 여러 연구가 보고 되었다. 제 2 장에서는 다양한 질병을 가진 환자들의 장내미생물을 이용하여 머신러닝기반 다중 분류 알고리즘으로 질병을 분류할 수 있는 모델을 구축하고 이를 평가하였다. 이 연구를 통해 LogitBoost 기반 예측 모델이 6 가지 질병을 가장 잘 구분 짓다는 것을 밝혔고, 미생물의 분류체계 중 속(genus)에서의 양을 이용했을 때 성능이 가장 좋다는 것을 보였다. 또한 미생물들을 선택하여 모델의 성능을 높이는 과정에서 다양한 질병을 동시에 구분하는 미생물들을 질병 진단을 위한 마커로 제시하였다.

인간에서 뿐만아니라 동물들에서도 장내미생물의 조성은 건강 및 생산량의 중요한 지표로 이용되고 있다. 예를들어, 사료에 따라 육계의 장내미생물 조성의 변화는 과거 연구에서 보고되어 왔다. 3 장에서는 삼채를 복용한 육계의 장내 미생물을 조사하고, 생산성과 연관이 있는 장내미생물들을 발굴하였다. 삼채의 잎을 복용은 육계의 장내미생물에 영향을 미치는 것을 밝혀내었으며, 연관성 분석을 통해서 삼채의 복용에 영향을 받는 미생물들이 육계의 체중, 경골강도 및 면역과 관련되어 있다 것을 제시하였다. 또한 미생물 기능분석을 통해 미생물 조성의 변화가 탄수화물 대사를 증진시킨다는 단서를 제시하였다.

정량적 데이터들의 좀 더 정확한 분석을 위해서는 정량화 단계에서 유전물질의 정확한 측정이 무엇보다 중요하다. 3 장에서는 전사체 분석에서 mRNA 의 발현량이 정확히 측정되지 못하게 하는 에러요인을 제시하고 이를 해결하기 위한 방법들을 제시하였다. 유전체 서열에서 mRNA 와 중첩되어 있는 lncRNA 는 정량화의 알고리즘상 mRNA 로 오인될 수 있다는 것을 가정하였으며 시뮬레이션을 통해 lncRNA 발현량임에도 mRNA 로 정량화되는 에러율을 제시하였다. 이러한 에러를 해결하기 위해서 정량화 단계에서 쓰이는 여러 알고리즘과 툴을 비교하여 더 정확한 정렬화를 통한 전사체 분석을 가능하도록 하였다.

주요어: 전사체 분석, 메타지놈 분석, 정량화, 머신러닝, lncRNA

학번: 2016-20462