Discovery of Population-Specific Translational Biomarkers among Korean Gastric Cancer Patients

Graduate School of Public Health
Yonsei University
Department of Biostatistics

Byeon Jae Hwan

Discovery of Population-Specific Translational Biomarkers among Korean Gastric Cancer Patients

Directed by Professor Chung Mo Nam

A Master's Thesis
Submitted to the Department of Biostatistics,
Division of the Graduate School Public Health
of Yonsei University
in partial fulfillment of the
requirements for the degree of
Master of Public Health

Byeon Jae Hwan

June 2020

This certifies that the Master's Thesis
of Jae Hwan Byeon is approved.

Thesis Committee Member : Chung Mo Nam

Thesis Committee Member : Sohee Park

Thesis Committee Member : Gyu Ri Kim

The Graduate School of Public Health
Yonsei University
June 2020

Acknowledgement

During the Master's Program in Health Information Statistics at Yonsei University's Graduate School of Public Health, I appreciate a lot of people to finish my thesis. I would like to thank you through this article.

Appreciate, Professor *Chungmo Nam*, the supervisor of Yonsei University's Health Information Statistics, who gave me a great direction and guided me to the end of my research. Thanks to the professor, I was very interested in the statistics and I was able to do good research. Thank you, Professor *Sohee Park*, who gave meticulous and careful review of papers and encouraged me while she was busy. Also, thank you, Professor *Gyuri Kim*, for getting interest in taking the thesis. Thanks a lot for my professors to many stories and advice in life, not just for study and research. Based on these teachings, I will continue to devote myself to academics and research and future.

It wasn't easy to work at the Macrogen Data Analysis Department and study at a graduate school, but I'd like to work hard because there was always a family next to me. Aunt *Seunghee Lee*, who gave advice like light whenever it was difficult, My mother, *Seungmi Lee* who always believed in and waited for support, and *Jaewoo Byeon*, who was my only younger sister and sometimes being friend and advisor. I really love all of my families and thanks.

While studying and researching, Macrogen's co-worker, *Sunghwan Bae* and *Shinsang Cho* helped with the analysis direction and research. Thanks to my Best friends in high school, who talked both bitter and sweet words to me, *Hyunbae Kim*, *Sunghwan Choi*, *Hyungkyu Kim*, and best friend in college, *Hyunsoo Chang*. Older brother *Bongsuk Cho* and *Hyukjin Kwon* who gave a lot of advice and help. Daebong Family, *Changhyun Park*, *Mingyu Lee*, and *Seunghoon Jeon*, who have always been at the company. Marc family *Kwangjun Lee, Eunsik Kim, Jiwan Kim, Namhee Kim*. Also, *Jaejun Lee*, *Bookyung Choi*, *YoomiSon*, and remember-family have always been great strengths as graduate students. Thanks to Koba family, J*inwoo Jeon, Joohyun Han, MinJun Kim, Woong Jung* for

hearing and considering my opinion for research. Lastly, I would like to express my gratitude to *Haneul Lee* who became the light when I was hard in my life.

I haven't mentioned all on the paper, but I would like to express my sincere gratitude to all those who have cared and encouraged me. I will do my best to become a man who is essential in the field of genetics and health by growing further and growing in the right form.

Thank you all for standing by my side throughout this journey. Thanks you.

Contents

List of Tables

List of Figures

Method. SRA and TCGA Data Analysis

Result. SRA and TCGA Data Analysis

Maching Learning

Analysis with selected gene

Annotation and Notwork

Abstract

Gastric cancer is one of the most commonly occurring cancers in the world, with an increased incidence and mortality, especially according to modern people's diet and busy lifestyle. In particular, the Koreans targeted in this study, gastric cancer occupies the leading cause of death in the entire cancer group and showing the record (2018. Statistics of causes of death). It is necessary to discover specific gastric cancer biomarkers for Koreans to predict the diagnosis of gastric cancer through the results of a study by the American Cancer Control Association (ACSI, 2009), in which the incidence and cause of different types of cancer differ to the biomarkers by comparing the difference in gene expression between the normal RNA group and the gastric cancer group.

NGS-based RNA-seq analysis was used as a method to see the difference in gene expression from RNA. Regarding RNA expression, analysis is performed using RNA expression data of TCGA database and Korean FASTQ data of NCBI SRA database to discover Korean gastric cancer biomarkers.

In the TCGA, data on gastric cancer was analyzed as RNA-seq expression data for 3 races (Asia, Caucasian, and Black). We were proceeding with the analysis of normal samples and type-specific gastric cancer samples, extracted differentially expressed genes in Asia. For SRA, FASTQ data from the Korean Gastric Cancer Project (Accession: PRJNA435914) were used and RNA-seq analysis wes performed. The RNA-seq pipeline wes being analyzed in the same method as the TCGA

RNA-seq pipeline, and normal samples and gastric cancer samples are analyzed in groups.

Genetic biomarkers for Korean-specific gastric cancer based on TCGA gastric cancer data by substituting it into the list of Korean differentially expressed genes to which a random forest model was applied for the Asian-differentiated expression genes selected by the AUC filter. Genes showing the statistically most significant value were confirmed in the three genes CIP2A, LDHD, and KIFC1. In addition, we were searching for g; profiler Annotation and GeneMANIA network for excavated genes, and derived the function of the pathway containing the Korean-specific genes.

Three TCGA-based Korean-specific genes extracted through this study (CIP2A, LDHD, KIFC1) and three g;profiler results (KEGG: 00260, GO: 0061846, GO: 0061845) and two GeneMANIA Network while hoping that the results will be utilized as a biomarker for the prediction of gastric cancer diagnosis in Koreans in the future.

Through the study of these RNA-seq, it was possible to recognize the difference between races in gastric cancer patients, in particular, gastric cancer appears to be more common in Asians than in whites and blacks. We expectd more and more safe and accurate biomarkers are researched for Asian and Korean.

I. Introduction

1.1. Gactric Cancer and RNA-seq

The stomach is the widest part of the digestive system and is located under the ribs of the left upper abdomen and is connected to the esophagus and duodenum. The stomach stores the ingested food, and the stored food is shredded into easy-to-digest forms through contraction and relaxation of gastric juice and secretion of gastric juice containing digestive fluid, and mixed to make it easier to digest. It serves to send food digested like porridge into the duodenum at an appropriate rate in accordance with the digestive and absorption functions of the small intestine. Stomach cancer is all malignant tumors that occur in the stomach, and generally refers to gastric cancer. Gastric cancer is a case in which adenocarcinoma located in the gastric mucosa causes malignant changes, accounting for 95% of malignant tumors occurring in the stomach. The gastric wall is composed of mucosal layer, submucosal layer, muscle layer, mesenteric layer, and subsidial layer. Stomach cancer starts from the gastric mucosa surrounding the inner side of the stomach and grows into the muscular layer and mesentery layer. Early detection and treatment are important because they invade the surrounding organs. (Severance Hospital Disease Information, 2014)

In this study, I intended to analyze through RNA-seq data of gastric cancer. Next-generation sequencing (NGS) is a more advanced gene and genome analysis technology than Sanger Sequencing, which is the first-generation sequencing technology used to diagnose gene mutations such as cancer and infectious diseases. Sanger sequencing has only one gene that can be identified at a time, but NGS can analyze all genes at once. (Macrogen press release, 2018)

RNA-seq is an analysis method that analyzes transcriptomes to identify differences in expression. Based on the central dogma, which becomes a protein through translation in transcript, it is a method of calculating by determining that the more number of transcripts, the greater expression. After transcriptome large-scale sequencing, a new one can be discovered through RNA-seq, and the expression value can be quantified. NGS-based RNA sequencing (RNA-seq) is considered to be the most powerful tool of modern medicine, beyond the limitations of existing methods for transcriptome analysis. RNA-seq is specialized in analyzing various types of non-coding RNA as well as changes in gene expression that has not been previously detected.

1.2. Needs to discover cancer markers by race

According to the mortality rate cancer death in 2018, which the Korean National Statistical Office surveyed, cancer was the most cause of death in Korea for 36 years, and the number of cancer deaths is the highest ever. According to the 2018 cause of death, the death rate from malignant neoplasms (cancer) last year (per 100,000 population) was 154.3 people. The death rate was the highest since statistics were recorded, increasing by 0.4 (0.2%) compared to 153.9 people last year. (Korean Statistics Office, 2018.)

In particular, as of 2018 year, gastric cancer was the fifth most common cancer worldwide, accounting for 8.2% of all cancer deaths and the third most common cause of death (Bray et al. 2018). Also, in Korea, there have been many changes in socio-economic, cultural and health aspects such as rapid aging, rapid economic growth, westernization of lifestyle, and the development of medical technology. (Go Gwang-pil, 2019). Gastric cancer was a trend that has declined globally in recent years (Torre et al. 2016), but according to data from the Central Cancer Registry in 2018, it was the cancer that ranks first among cancers in Korea out of thyroid cancer in

2016. In 2016, the number of gastric cancer patients was 30,504, and the age-standardized gastric cancer incidence rate was 35.4 prople per 100,000 population (Korea Central Cancer Register, 2018).

According to the Journal of Cancer Control Society in USA, Koreans showed the biggest difference from the American cancer group, especially the Caucasian group. Korean males had an incidence of stomach cancer, liver cancer, and gallbladder cancer up to 4.4 times higher than in whites. Based on white males, the incidence of gastric cancer among Korean males was 9.83, which was 10 times higher. On the other hand, in other types of cancer, such as esophageal cancer, laryngeal cancer, and pancreatic cancer, the incidence rate of Korean males was lower than that of Caucasians, and Korean females also showed significantly higher incidences of gastric cancer (4.52) and liver cancer (4.48) than Caucasians. Korean women are significantly higher than Caucasians with gastric cancer 8.04, liver cancer, 6.37, gallbladder cancer 3.52, and cervical cancer 2.48.

Through the study of the current status of cancer and cancer mortality and survival rate among races (American Cancer Society ACS, 2009), it is necessary to find biomarkers for each race according to each cancer type. It is derived through gene expression analysis.

II. Method & Materials

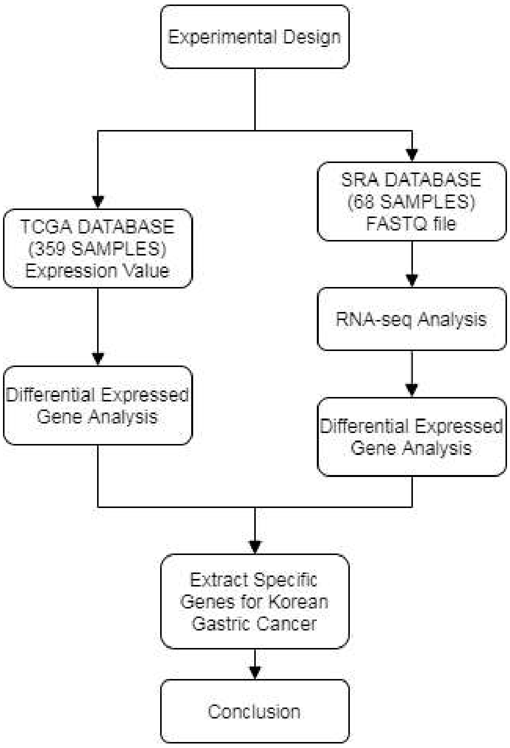

Experimental Design
TCGA DATABASE (359 SAMPLES) Expression Value
SRA DATABASE (68 SAMPLES) FASTQ file
RNA-seq Analysis
Differential Expressed Gene Analysis
Differential Expressed Gene Analysis
Extract Specific Genes for Korean Gastric Cancer
Conclusion

Figure 1. Overall Workflow

2.1. TCGA and NCBI SRA database

This study was based on data from The Cancer Genome Atlas (TCGA) and Sequence Read Archive (SRA). The TCGA is a data portal that collects all cancer data from cancer-related projects around the world and analyzes them using its own standardized analysis method to produce, manage, and provide data. The SRA is a database managed by the National Center for Biotechnology and Information Technology (NCBI). It is a database that stores and manages sequence read data from various projects.

Analysis was performed based on RNA-seq profiling (Read-Count type) data of samples related to gastric cancer in the TCGA Portal. To see differences by race, only three races (White, Black, Asian) were filtered and analyzed with a total of 359 sample data. In the SRA database, 68 Korean gastric cancer data (FASTQ type) corresponding to SRX3763198 to SRX2763265 were provided and analyzed. Both databases are extracted from samples of patients with gastric cancer, and their age ranges from 60 to 70 years old.

Table 1. Data Source and Samples

Resource Name	Number of Samples
SRA Experiments (SRX3763198 ~ SRX3763265)	68
- Normal	34
- Cancer	34
TCGA Experiments (Gastric Cancer)	359
- Solid Tissue Normal	23
Asian	5
Black	1
White	17
- Tumor Tissue	336
Asian	64
Black	11
White	261

2.2. edgeR in R package

2.2.1. TMM Normalization

In the process of using RPKM used from Microarray for RNAseq, as incompatibility is observed, various normalization methods have been studied. The RLE and edgeR methods are studied considering the division of the total mapped reads do not reflect the entire transcriptome size (Wangner, G. P. et al., 2012). When transcriptome size is defined as a factor affecting the observed total read, it can be divided into RLE and TMM methods according to the method of correcting this size factor. First, RLE is a method of calculating geometric mean and median value and estimating it as a size factor to obtain normalized values (Anders et al. 2010). TMM is a method using the weighted mean after excluding genes with large log ratio and expression level (Robinson, M. D. & Oshlack, A. 2010). There are some papers that show better performance by comparing these two methods with RPKM (Wagner, GP, 2012), but there are some aspects that it is better to use RPKM that considers the gene length. As research is conducted, it is recommended to study using Read-Count without considering the gene length.

edgeR Filter: Only genes with p-values below the specified value are selected using the edgeR analysis results. Generally, a value of 0.05 or less is used, genes having a p-value of 0.05 or less are selected. In this study, analysis is performed using a value of adjust p-value (Benjamini hochberg) of 0.01 or less.

edgeR is an R package that can be used in similar environments such as DESeq, and is relatively fast in terms of speed like DESeq, also can be analyzed when there are other factors besides the controlled conditions. And since Bayesian gene-wise dispersion estimation can be calculated, information between genes/transcripts can be obtained. However, if the replicate is not considered, the common and tag-wise dispersion cannot be obtained. In this case, the following two methods are used in edgeR. The first method is to assume that the biological coefficient of variation (BCV) used by edgeR. It is assumed to be some empirically known constant according to the data, which may deviate from the original analysis direction, and the dispersion is larger than actually expected. However, it is much more realistic in that analyzes with dispersion rather than excluding and analyzing biological variation. The second method is to estimate dispersion from the control transcript, for example, using housekeeping gene. Since the housekeeping gene is having a small change in expression level among cells, it can be assumed that it is suitable as a common dispersion.

2.2.2. Fisher's Exact Test for DEG

Fisher exact test is a statistical method used to perform gene-enrichment analysis. The Fisher exact test is for the hypothesis that gene sets and genomes have the same ratio that can be divided into two categories: In Pathway and Not in Pathway.

When the p-value of Fisher exact test of a pathway A is calculated, according to the decision, when the p-value is less than the significance level under the significance level of 0.05, the ratio of the pathway that the researcher's gene set comprises is the total genome. It can be said that it is different from the proportion of pathways that are formed. In other words, a specific pathway is more relevant (associated, enriched) to the gene set of interest.

In DEG analysis and pathway analysis, more than 4,000 hypothesis tests (more than 4,000 p-values) are performed at the same time. In other words, the p-value that can be determined by gene or pathway is calculated. Here, researchers face multiple testing problems. This is because too many false positives are calculated, and when there are genes that do not actually differ, the probability of determining that there is a difference increases. For example, the probability of making the right decision under significance level 0.05 (5%) is 0.95 (95%). At this time, if the number of genes to be determined is 10,000, the probability of making the correct judgment for all genes is reduced as “0.95 X 0.95 X …”. Therefore, it is necessary to calculate the p-value according to the significance level to be lowered to make the right decision. This is called a multiple testing problem. To solve this problem, methods for correcting p-values such as Benjamini-Hochberg, Bonferroni correction, and permutation test have been studied, and DEG or pathway analysis packages provide these corrected p-values.

The proposed method considering the ratio of true positive and false positive is a correction method from the perspective of FDR. One of the methods is proposed by Benjamini-Hochberg. It is possible to find the corrected values in order from the smallest p-value. This is a correction method in which the false positive is reduced relatively gently.

2.3 ROC Curve

The ROC (receiver operating characteristic) curve is one of the traditionally used methods for evaluating the accuracy of discrimination in the field of discriminant analysis (Kang et al., 2014; Jeon, Lee, 2014). After calculating the sensitivity and specificity for each cut value of the predictor, a ROC curve is created by drawing the connecting line with the corresponding and sensitivity as the vertical axis and the '1-specificity' as the horizontal axis (Simundic, 2009).

Area under curve (AUC) means the area under the ROC curve. The good model has better performance, the closer ROC curve is to the upper left of the diagonal. Also, the diagonal means the ROC curve by chance (that is, the diagonal means the ROC curve for a model with no predictive power). Using these properties, AUC can be used as a measure the performance of a model. If the ROC curve coincides with the diagonal, AUC = 0.5, so AUC has a value between 0.5 and 1, and the closer to 1, the better the performance of the model.

When analyzing a single data, it is generally desirable to analyze as many models as possible (Nam et al., 2017; Ryu, Hwang, 2017). Therefore, in order to obtain an optimal model, it is necessary to compare and evaluate several models, and if one model is selected, it must be proved that the selected model is superior to other models. A variety of case studies show that AUC can be useful in evaluating the performance of variables or models in real problems. Also, the ROC curve can be used efficiently in the problem of determining the optimal truncation value.

The AUC and pROC package in R allow the creation of ROC curves and the calculation of AUC, the LOGISTIC Procedure and pROC package allow comparison and testing of multiple AUCs.

2.4 Random Forest

Random Forest is a model proposed by Breiman (2001) by applying an ensemble technique to solve the overfitting problem in the decision tree model. In other words, it is a method of generating a large number of bootstrap samples and applying a decision tree model to synthesize the results. The main difference from other ensemble models is not only randomness is introduced in the part where the bootstrap sample is generated, but randomness is also introduced when selecting explanatory variables at each node when the decision tree model is fitted (Eugene, 2015). Accordingly, the random forest has maximum randomness, which reduces the correlation between decision trees and reduces prediction errors. In addition, as the number of decision trees increases, the prediction error decreases, and even if the number of decision trees increases, this model has the advantage of not being overfit.

In general, the model verification method divides the data into training data and verification data. Use the method to find the static classification rate by verifying by pressing or cross-validation. On the other hand, in random forest, there is no need to divide data into training and verification data. When creating this bootstrap sample, the data not selected as the bootstrap sample are called out-of-bag data (OBB) and can be used to perform model verification by using it instead of the verification data.

Random Forest uses only one variable when branching the existing decision tree. The disadvantages of the model's high explanatory power but poor predictive power and poor model stability. It is solved by bootstrapping with maximum randomness, and it provides a highly stable model with high predictive power, especially when there are many explanatory variables. In addition, random forest is a single method using existing parameters such as maximum likelihood method. More accurate and better than machine learning algorithms such as decision trees and neural networks is emerging as an alternative to produce results, and there are many research data available, especially in gene expression studies.

2.5. RNA-seq Analysis Pipeline

TCGA data provides RNA expression results in the Database Portal. Therefore, TCGA performs DEG analysis (differential expression gene analysis) without separate RNA-seq analysis.

In the case of NCBI SRA database, raw data was provided as a FASTQ file. Therefore, in order to calculate RNA expression, the pre-processing

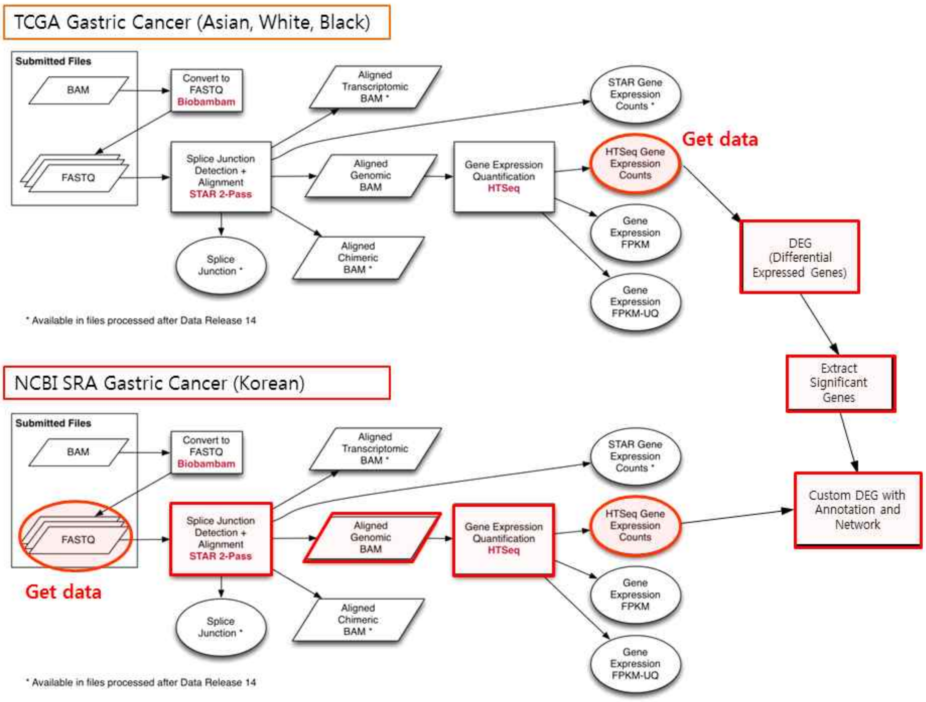

TCGA Gastric Cancer (Asian, White, Black)
Submitted Files
BAM
FASTQ
Convert to FASTQ Biobambam
Aligned Transcriptomic BAM *
STAR Gene Expression Counts *
Splice Junction Detection + Alignment STAR 2-Pass
Aligned Genomic BAM
Gene Expression Quantification HTSeq
HTSeq Gene Expression Counts
Get data
Gene Expression FPKM
Gene Expression FPKM-UQ
Splice Junction *
Aligned Chimeric BAM *
* Available in files processed after Data Release 14
DEG (Differential Expressed Genes)
Extract Significant Genes
NCBI SRA Gastric Cancer (Korean)
Submitted Files
BAM
FASTQ
Get data
Convert to FASTQ Biobambam
Aligned Transcriptomic BAM *
STAR Gene Expression Counts *
Splice Junction Detection + Alignment STAR 2-Pass
Aligned Genomic BAM
Gene Expression Quantification HTSeq
HTSeq Gene Expression Counts
Custom DEG with Annotation and Network
Gene Expression FPKM
Gene Expression FPKM-UQ
Splice Junction *
Aligned Chimeric BAM *
* Available in files processed after Data Release 14

Figure 2. RNA-seq Analysis Pipeline (reference; TCGA pipeline)

process of FASTQ file trimming, mapping to the reference genome, and profiling analysis were performed. The analysis tools were conducted in the STAR-HTseq pipeline in line with TCGA's analysis tools, and the reference genome and annotation files were used GRCh38.d1.vd1 version.

As a result of DEG analysis using Expression Profiling data of the two databases, ROC curve and Random Forest analysis were performed to extract 15 Korean-specific genes. Attach the description of the RNA-seq analysis tools below. The analysis proceeds to Python 2.7 and R 3.6.3 versions. Trimmomatic 0.32 ver. is a program that performs trimming depending on various parameters on illumina paired-end or single-end. STAR 2.6.0c, Spliced Transcripts Alignment to a Reference (STAR) software based on RNA-seq alignment algorithm which utilizes sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure.[4] HTseq version 0.10.0 is a Python package that provides infrastructure to process data from high-throughput sequencing assays.[7] A very typical use case for the HTSeq library is to for a given list of genomic features (such as genes, exons, ChIP-Seq peaks, or the like), how many sequencing reads overlap each of the features.

III. Method; SRA database

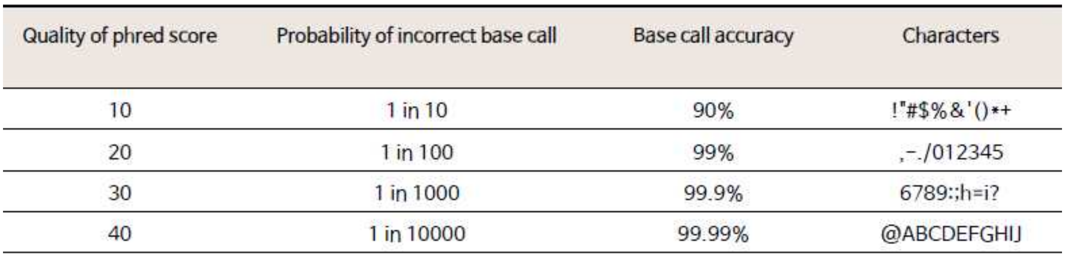
3.1 RNA-seq Analysis Summary

Differential expression genes were analyzed by obtaining gene expression values through transcriptome sequencing of Homo sapiens (GRCh38). Pre-processed trimmed reads are mapped to a known reference genome using a STAR program capable of splice junction processing.

Sequencing of paired-end transcripts for a total of 68 samples from NCBI SRA, Korean Gastric Cancer, results were produced within the normal range of all samples. The raw data for each sample and the trimmed read after the pre-processing process were compared with the total data volume and Q30 (phred score, base quality of 30% or more %), respectively. The data (Table 6, Table 7) after this paper could be confirmed.

The Fred quality score was a numerical representation of how accurate each base is, and the higher Q number, the higher accuracy of that base. Q20 has a wrong base probability of 1% and Q30 has a wrong base probability of 0.1%. The Fred quality score Q was calculated as -10log10P, where P represents the probability of a base call error.

Table 2. Quality of Phred Score



Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

The pre-processed trimmed reads were mapped to a known reference genome using a STAR program capable of splice junction processing. The mapping ratio defined by mapped reads versus the number of trimmed reads for each sample can be confirmed in the table after this paper.

After read mapping, used the HTseq program to extract the read count for each gene of each sample based on the gene annotation of the species.

Using this value, DEG (Differentially Expressed Genes) analysis was performed using edgeR for the comparative combination (Test_vs_Control), and |fc|>=3 & exactTest adj.p<0.01 to select genes expressing differentially between the two groups.

3.2. Pretreatment analysis of differential expressed genes

3.2.1. Analysis data quality check and pre-processing

With the count value for the known gene obtained as a result of read mapping, the process of selecting the differentially expressed gene between samples was performed. In the pre-processing process, quality check of data and normalization between samples were performed before entering the analysis, and similarity between samples was checked when biological replicates exist to determine whether data was reliable.

3.2.2. Data Quality Check

For each gene, genes having a count value of 0 in at least one sample from a total of 68 samples were excluded from analysis. Therefore, statistical analysis was performed on 7,511 genes excluding 45,461 genes out of 52,972 genes.

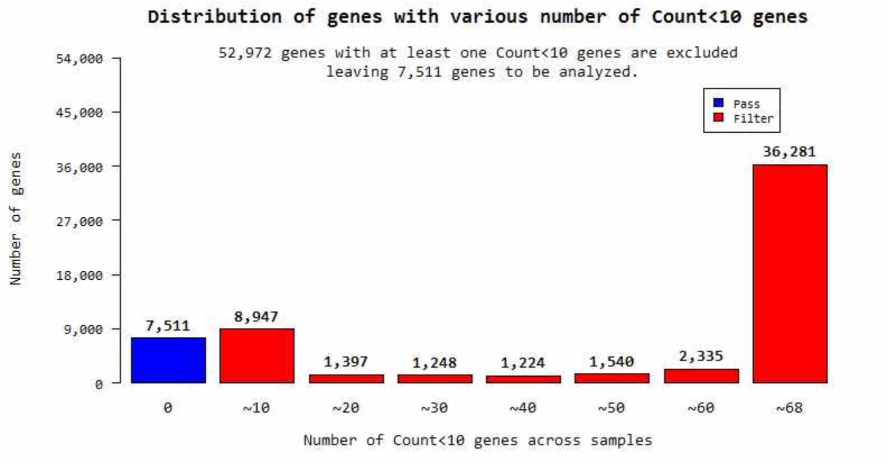


Figure 3. SRA Distribution of genes for analysis

3.2.3. Data transformation and normalization

To reduce systematic bias that could affect biological meaning in comparison between samples, the size factor was estimated using count data, and the data was corrected using Trimmed Mean of M-values (TMM) normalization. (using the'edgeR' R library).

3.2.3.1. Box plot of distribution of expression values per sample

Figure 4. is the percentile, median, percentile to show the distribution of expression values for each sample for before/after Logarithm (based 2) of raw signal(Count)+1 and TMM normalization. Bar plot visually expressed using the minimum value. Logarithm was performed by adding 1 to the read count value to see the value when the log of the low unit is taken.

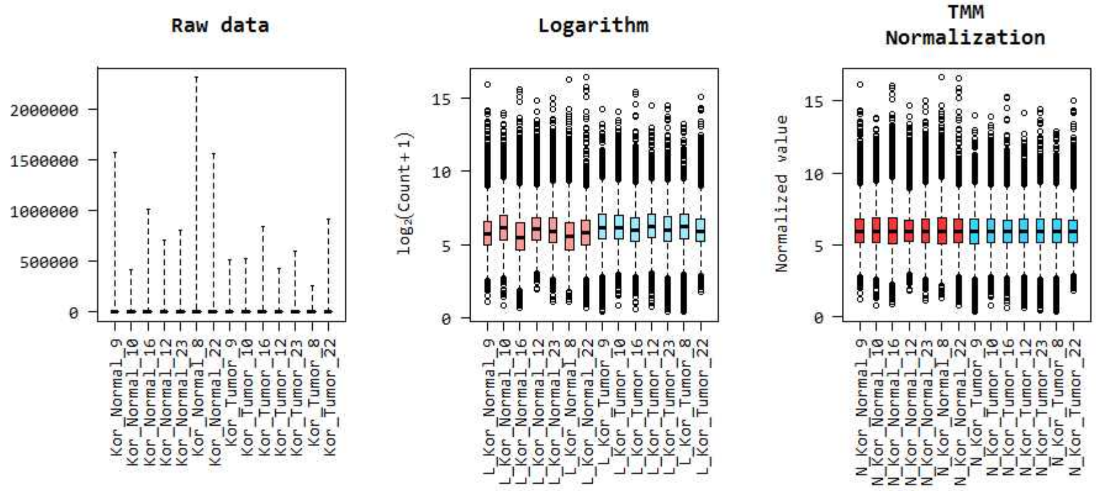


Figure 4. SRA Distribution after pre-processing

3.2.4. Correlation analysis between samples

We used the Log2(Count+1) value to check the degree of similarity between samples (Pearson′s coefficient, Pearson′s correlation coefficient) to check whether the repeat sample is reproducible. (Range: $-1 \leq r \leq 1$) The closer the correlation coefficient value was to 1, the higher the similarity between samples. Overall, since the blue coefficient was high, it could be confirmed that the similarity between SRA samples was high.

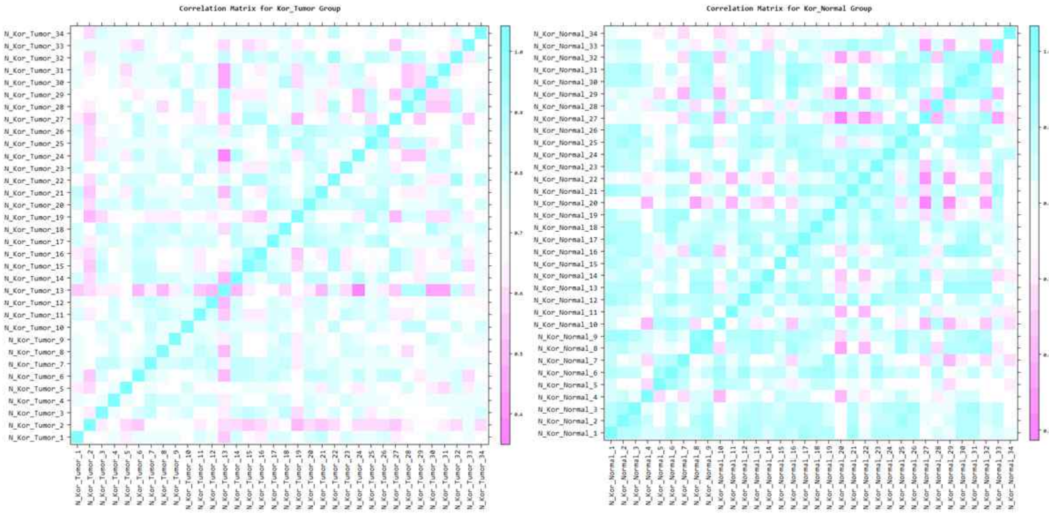


Figure 5. SRA Tumor, Normal Samples level plot

3.2.5. MDS (Multidimensional Scaling)

MDS is a picture expressed in two-dimensional space by using two components that best describe the degree of similarity between samples using the Log2(Count+1) value. It can be confirmed that there are outlier samples and similar expression patterns between biological replicates.

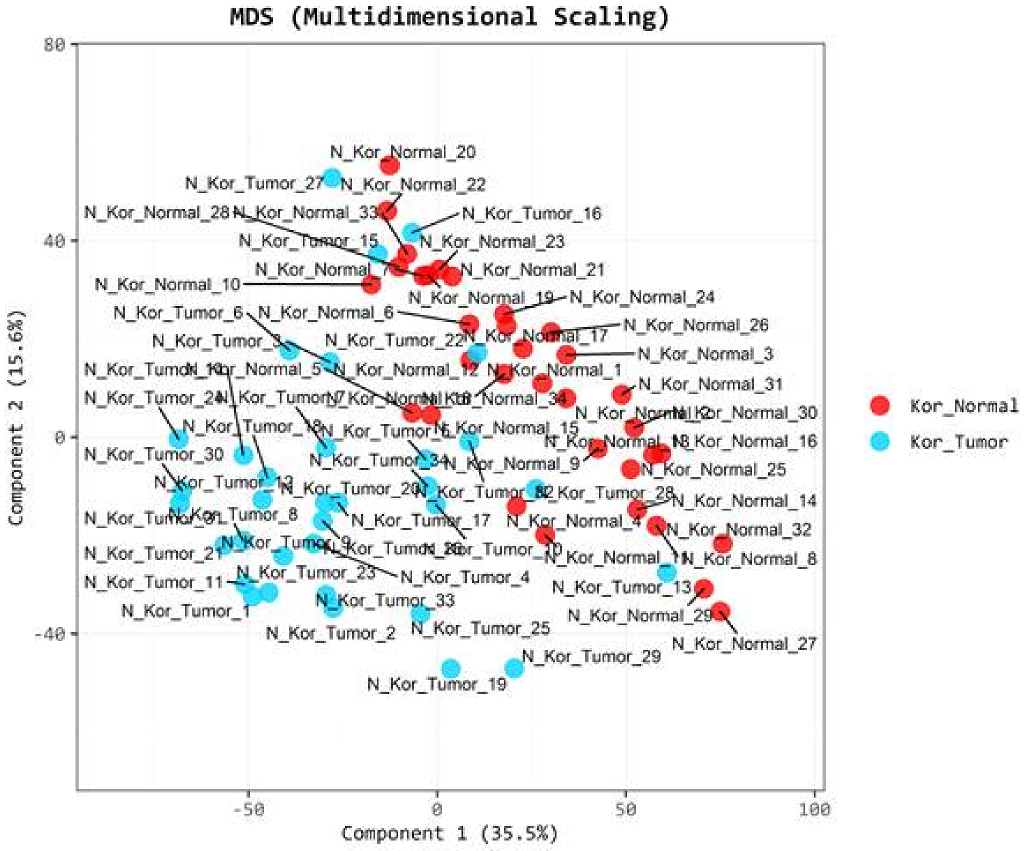


Figure 6. SRA Multidimensional Scaling (MDS)

IV. Method; TCGA Database

4.1. Pretreatment analysis of differential expressed genes

Genes with a count value of 0 in at least one sample from all 359 samples were excluded from the analysis. Therefore, statistical analysis was performed on 12,562 genes excluding 35,359 genes out of a total of 47,921 genes.

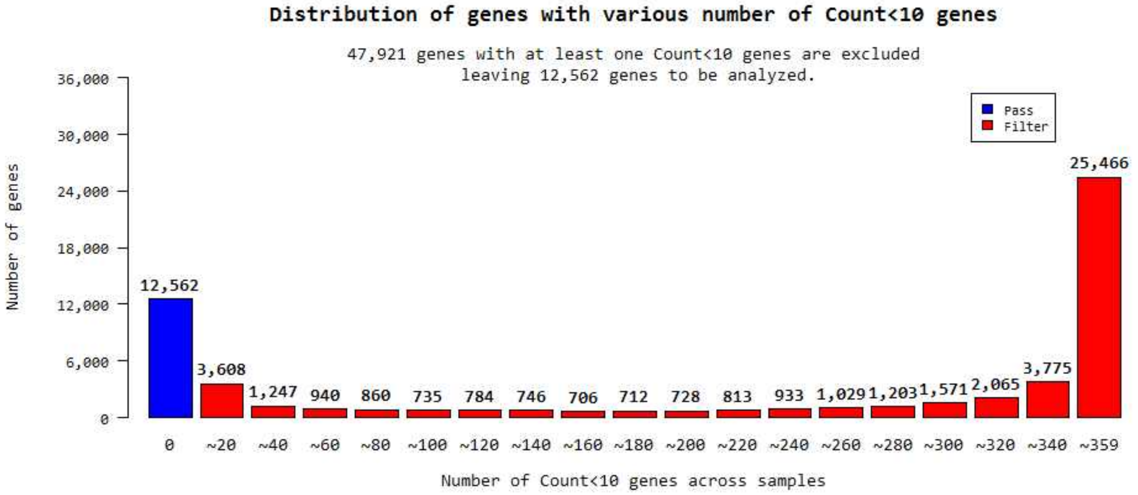


Figure 7. TCGA Distribution of genes for analysis

4.1.1. Data transformation and normalization

To reduce systematic bias that can affect biological meaning in comparison between samples, the size factor was estimated using count data, and the data was corrected using Trimmed Mean of M-values (TMM) normalization. (using the ‘edgeR’ R library).

4.1.2. Box plot of distribution of expression values per samp le

Figure 8. is the percentile, median, percentile to show the distribution of expression values for each sample for before/after Logarithm (based 2) of raw signal(Count)+1 and TMM normalization. Bar plot visually expressed using the minimum value.

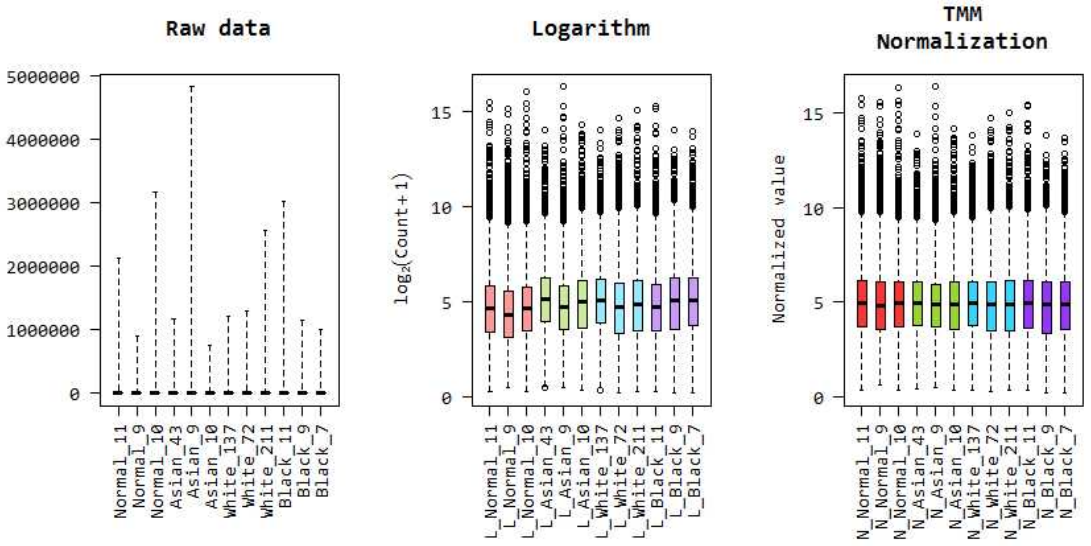


Figure 8. TCGA Box plot for Distribution after pre-processing

4.1.3. Correlation analysis between samples

We used the Log2(Count+1) value to check the similarity (Pearson's coefficient, Pearson's correlation coefficient) between samples to check the repeatability of repetitive samples. (Range: $-1 \leq r \leq 1$) The closer the correlation coefficient value was to 1, the higher the similarity between samples. Samples were reproducible in the White race, Asian and Normal groups, but it was confirmed that there are differences between the samples in the Black race.

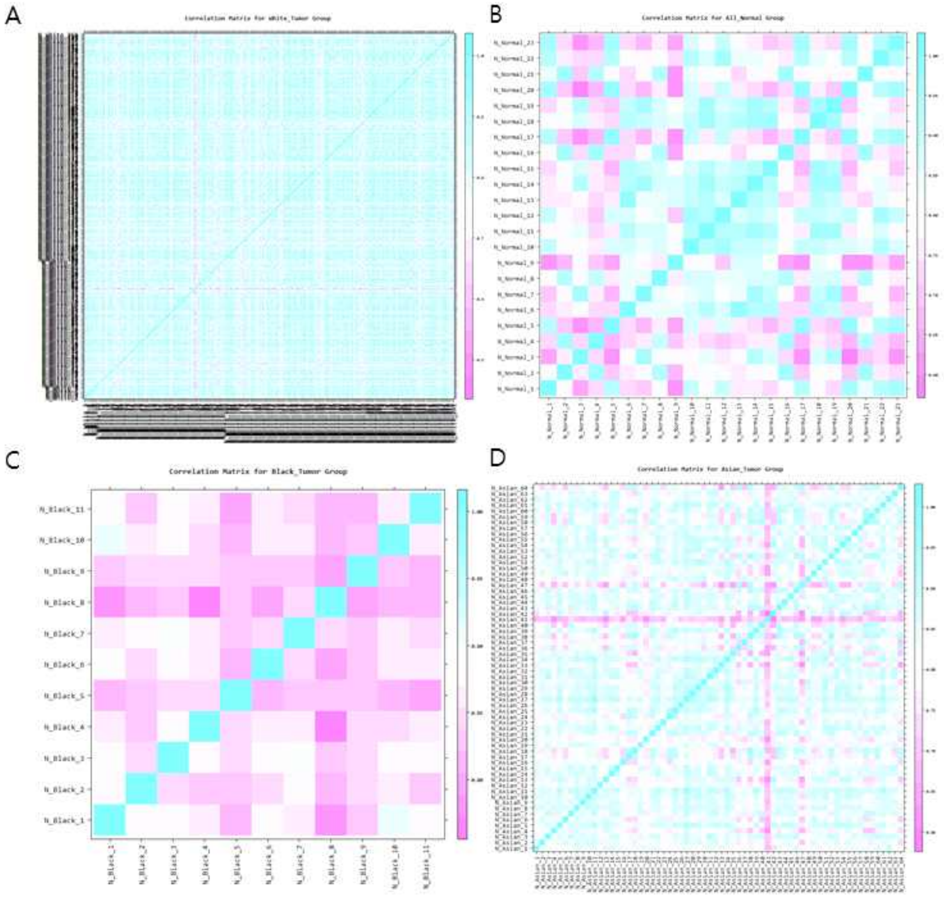


Figure 9. TCGA level plot. (A) White, (B) Normal (C) Black (D) Asian

4.1.4. MDS, Multidimensional Scaling

MDS is a picture expressed in two-dimensional space by using two components that best describe the degree of similarity between samples by using Log2(Count+1) values. It can be confirmed that there are outlier samples and similar expression patterns between biological replicates. It can be seen that the samples are distributed in a certain pattern in the 4 groups.

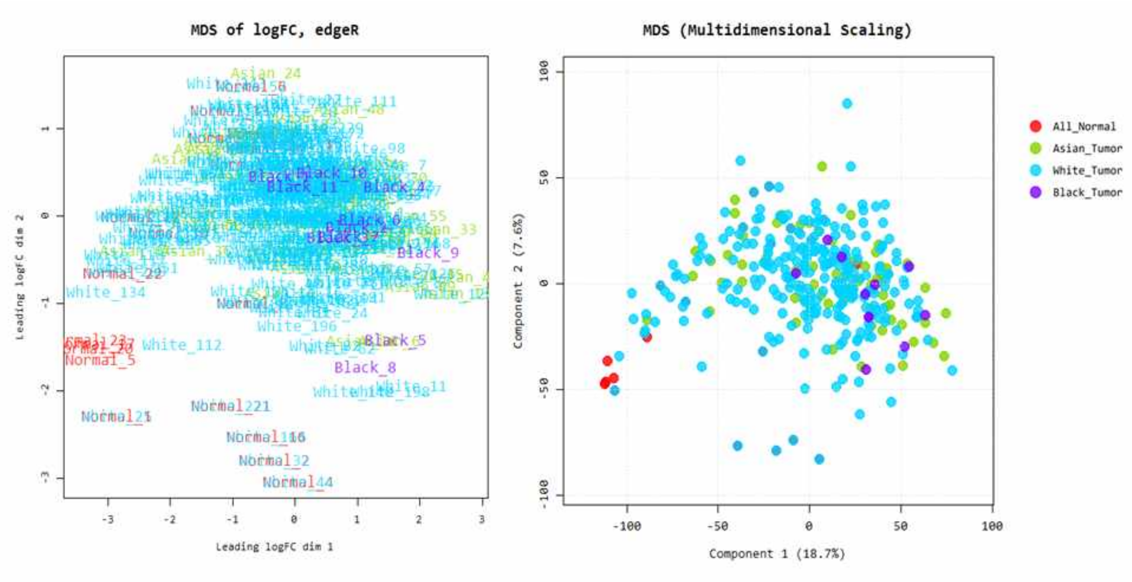


Figure 10. TCGA Multidimensional Scaling (MDS)

V. Result

5.1. SRA Analysis result of differential expressed genes

The sequence of DEG (Differentially Expressed Genes) analysis. Original Raw Data was targeted to count values for known genes obtained through HTseq.

After filtering genes with low quality during data pre-processing and QC, TMM normalization was performed.

For statistical analysis, |fc|>=3 & exactTest adj.p<0.01 for each comparison combination was used.

The level of similarity for each gene was grouped by hierarchical clustering analysis on the list of significant genes, and this was visualized by heatmap and dendrogram.

5.1.1. Number of Genes per Up and Down based on Fold Change and p-value

It shows the number of significant genes based on fold change and p-value of the Tumor vs Normal comparison combination.

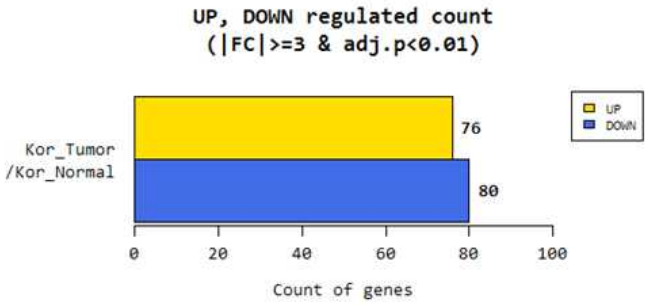


Figure 11. SRA UP, DOWN regulated count

5.1.2. Volcano plot of expression values between two groups

This is a plot showing the log-fold change of the expression value between the Tumor vs Normal comparison combination and the p-value derived through comparison between the two groups as a volcano plot. (X axis: log2 Fold change, Y axis: -log10 p-value)

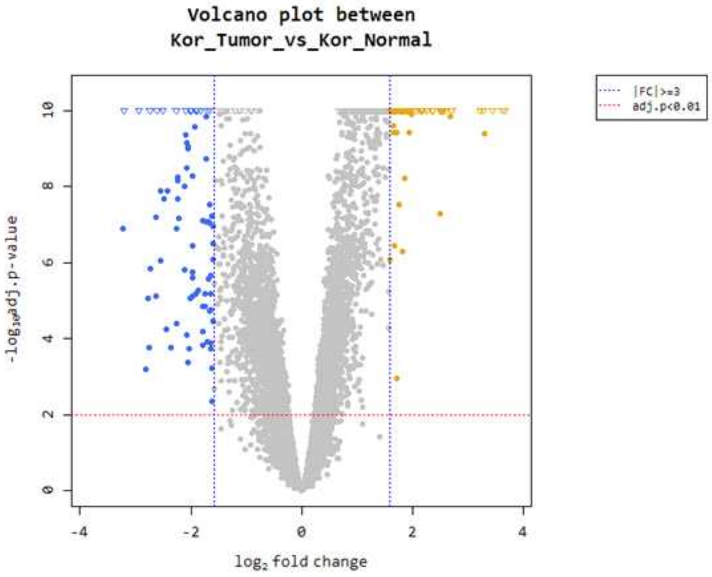


Figure 12. SRA Volcano Plot

5.1.3. Differential expression gene expression according to expression intensity, Smear Plot

Smear plots (X-axis: average log2 counts-per-million (logCPM), Y-axis: log2 Fold Change) are used to identify genes in the test group compared to the control with high average of the expression values of the two groups.

Even if the change in fold is more than 2 times the same, a gene with a difference at a higher place may have higher reliability than a place where the average expression value is 2 times or more different.

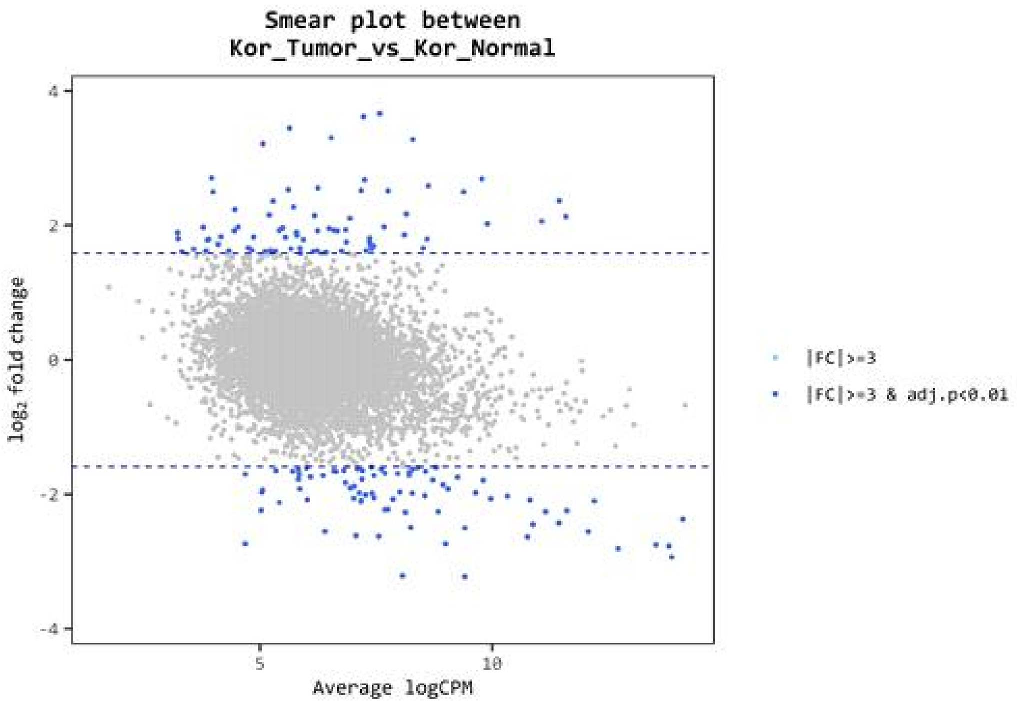


Figure 13. SRA Smear Plot

5.1.4. Hierarchical Clustering Analysis

For the significant DEG list, samples and genes with similar expression levels are grouped and displayed through hierarchical clustering analysis (Euclidean Distance, Complete Linkage) using the normalized values of each sample. When the heatmap is drawn with significant genes by comparing the Normal group and the Tumor group, it can be seen that the pattern is divided according to the Normal and Tumor groups.

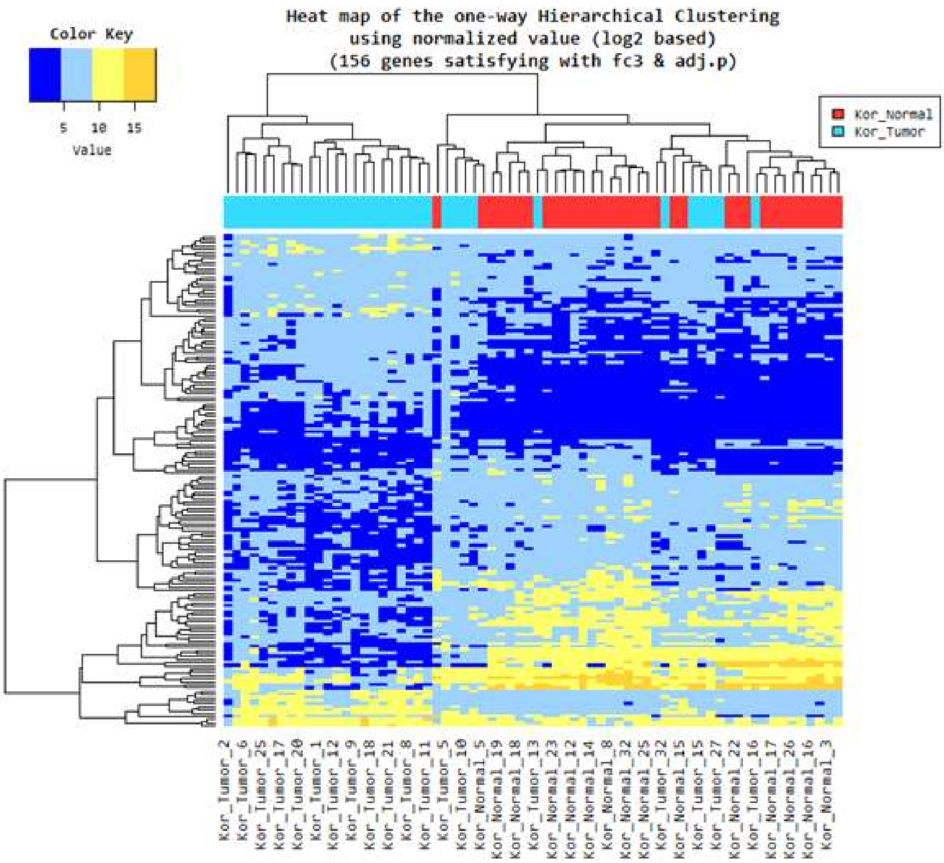


Figure 14. SRA Heatmap

5.2 TCGA Analysis result of differential expressed genes

The following description shows the order of DEG (Differentially Expressed Genes) analysis. Original Raw Data is targeted to count values for known genes obtained through HTseq. After filtering genes with low quality during data pre-processing and QC, TMM normalization is performed. For statistical analysis, |fc|>=3 & exactTest adj.p<0.01 for each comparison combination. Significant results are selected with |fc|>=3 & exactTest adj.p<0.01. The level of similarity for each gene is grouped by hierarchical clustering analysis on the list of significant genes, and visualized by heatmap and dendrogram.

5.2.1. Number of genes per Up and Down based on the Fold Change and p-value

Shows the number of significant genes based on fold change and p-value for each comparison combination.

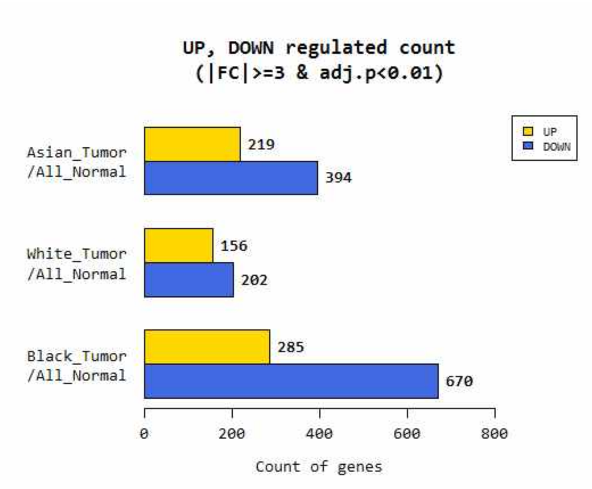


Figure 15. TCGA UP, DOWN regulated count

5.2.2. Hierarchical Clustering Analysis

For the significant DEG list, samples and genes with similar expression levels are grouped and expressed through hierarchical clustering analysis (Euclidean Distance, Complete Linkage) using the normalized values of each sample.

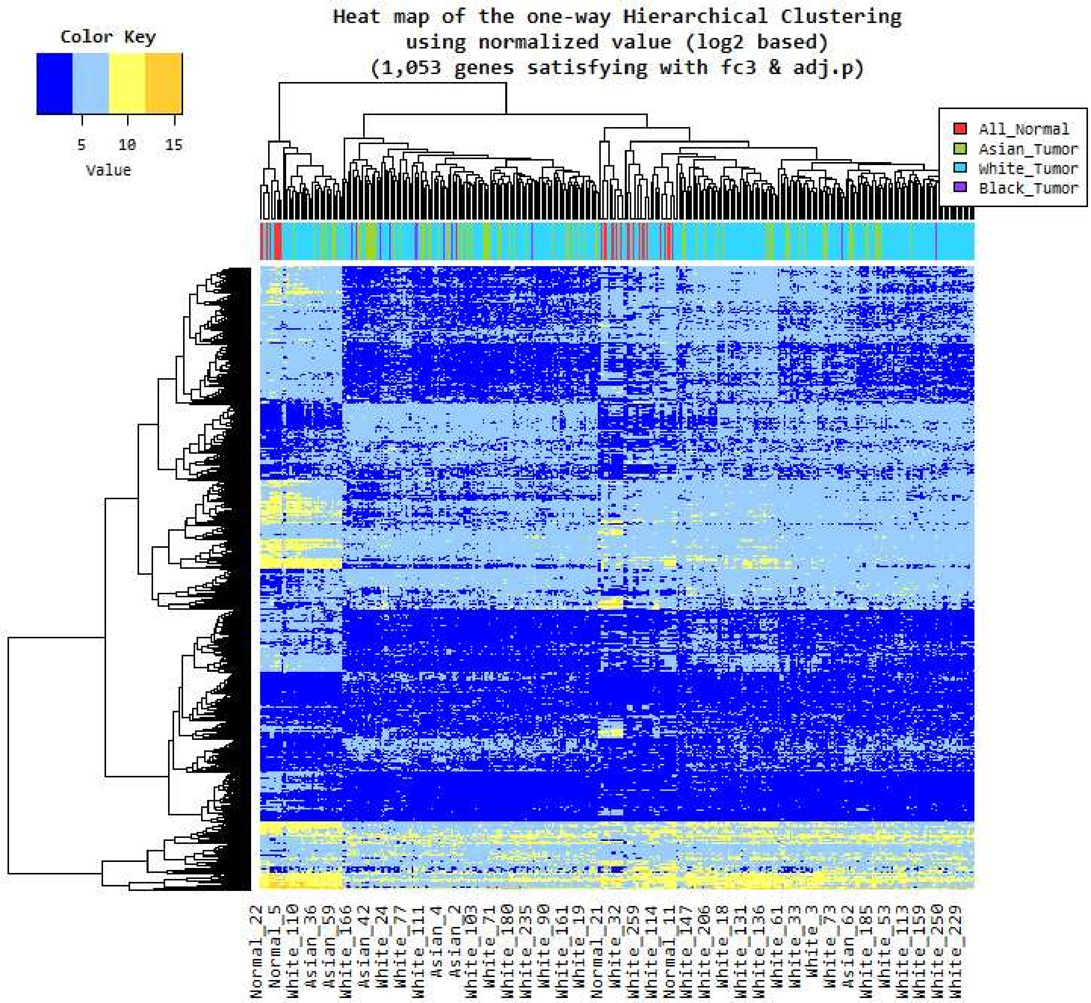


Figure 16. TCGA Heatmap

`

5.2.3. Venn-Diagram with DEGs

Genes obtained by conducting DEG analysis of each Tumor group (Asian, Black, White) compared to the Normal group are drawn with a Venn Diagram to distinguish between common and non-genetic genes.

It is possible to confirm that genes with different differences exist even in the same stomach cancer group for each race. The purpose of this study is to extract specialized genes from SRA Korean data, and Koreans belong to Asian races. As shown in Figure 17. (Asian_Tumor vs. All_Normal) in Venn Diagram, genes that showed significance in Asian gastric cancer groups compared to all normal groups are analyzed. Proceeded.

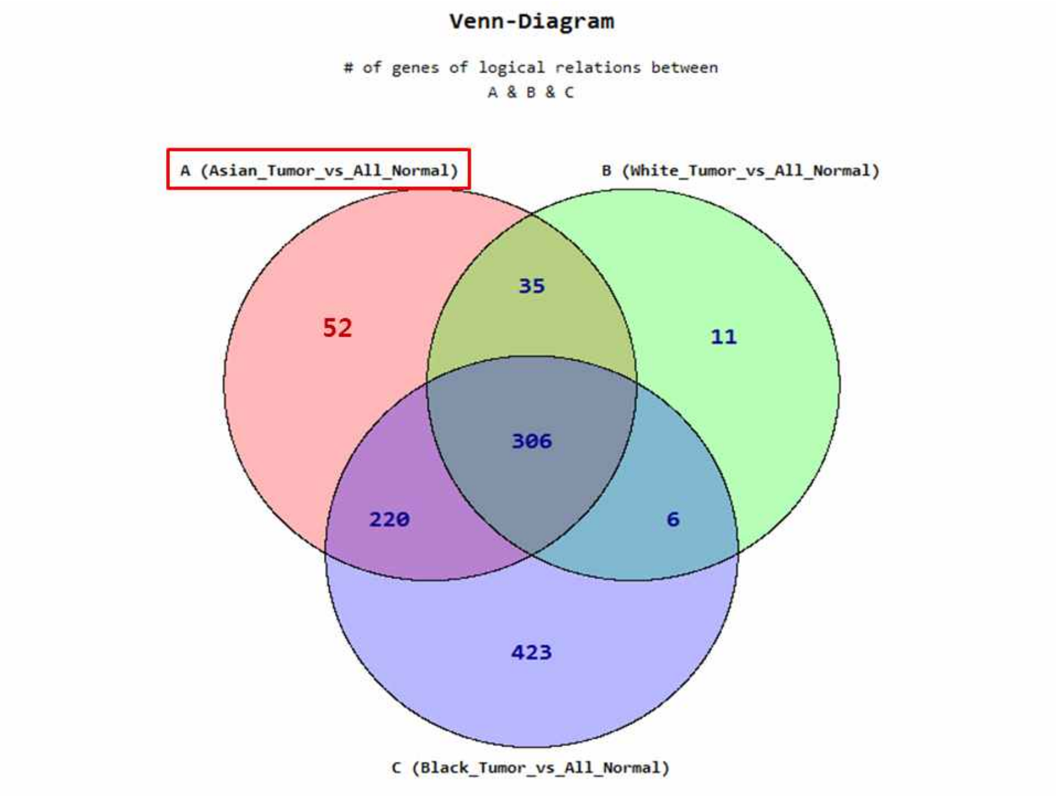


Figure 17. TCGA Venn Diagram

5.3 Extract Korean Specific genes for Gastric Cancer Data

5.3.1 TCGA Asian ROC curve

Through the TCGA gastric cancer DEG analysis, 52 DEG gene listed from the Asian group are extracted, and the AUC values of the genes were respectively obtained, and 28 genes with 0.8 or higher are extracted. (Table 2)

Table 3. Gene list of AUC result (over 0.8)

Gene	AUC	Description
NDRG2	0.991847826	NDRG family member 2
CFAP157	0.953804348	cilia and flagella associated protein 157
ATAD5	0.953804348	ATPase family AAA domain containing 5
CGAS	0.940896739	cyclic GMP-AMP synthase
CIP2A	0.938858696	cell proliferation regulating inhibitor of protein
DLEU2	0.918478261	deleted in lymphocytic leukemia 2
BBC3	0.911684783	BCL2 binding component 3
TYMS	0.901494565	thymidylate synthetase
CCDC34	0.899456522	coiled-coil domain containing 34
ANKRD10-IT1	0.893342391	ANKRD10 intronic transcript1
TPM3P9	0.889266304	tropomyosin 3 pseudogene 9
KIFC1	0.887907609	kinesin family member C1
ME1	0.886548913	malic enzyme 1
BRIP1	0.878396739	BRCA1 interacting protein C-terminal helicase 1
CKAP2L	0.866168478	cytoskeleton associated protein 2 like
SCARA3	0.864130435	scavenger receptor class A member 3
AL591895.1	0.861413043	novel transcript
MPP7	0.861413043	membrane palmitoylated protein 7
BAIAP2	0.849184783	BAR/IMD domain containing adaptor protein 2
ANGPTL4	0.839673913	angiopoietin like 4
AATF	0.835597826	apoptosis antagonizing transcription factor
OSGIN1	0.832880435	oxidative stress induced growth inhibitor 1
PID1	0.828125	phosphotyrosine interaction domain containing 1
PVR	0.827445652	PVR cell adhesion molecule
CLIP4	0.826766304	CAP-Gly domain containing linker protein family
LDHD	0.809103261	lactate dehydrogenase D
CTF1	0.80638587	cardiotrophin 1
ZNF853	0.805027174	zinc finger protein 853

Among the genes with AUC 0.8 or higher, the ROC curve was drawn as the Top 6 Gene. (Figure 18). The Curve was drawn with 1-Specificity on the X-axis and sensitivity on the Y-axis, and it can be seen as a Gene with higher predictive power as the area under the curve is closer to 1.

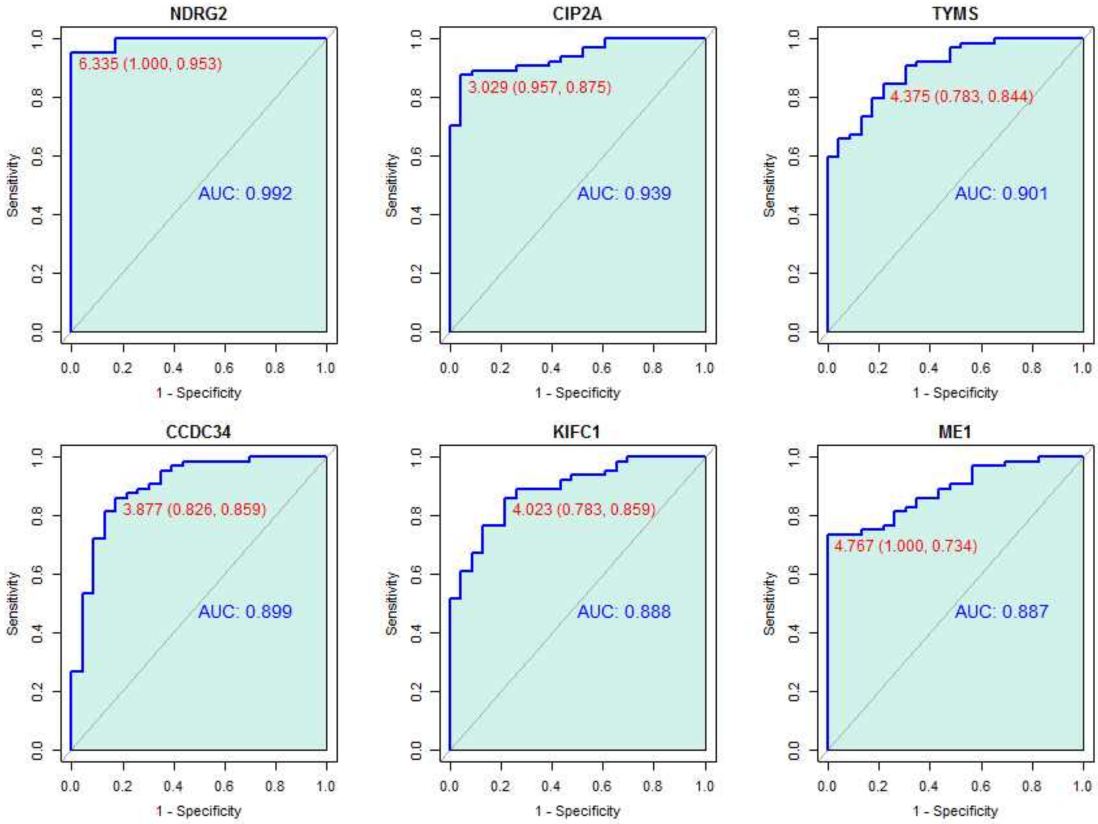


Figure 18. ROC Curve for top 6 genes

5.1.2 Good fit of predicted genes through the Random Forest

After setting the variable that divided the samples into Normal and Tumor groups for genes with AUC 0.8 or higher, and applying them to the Random Forest model 100 times, OOB error is obtained.

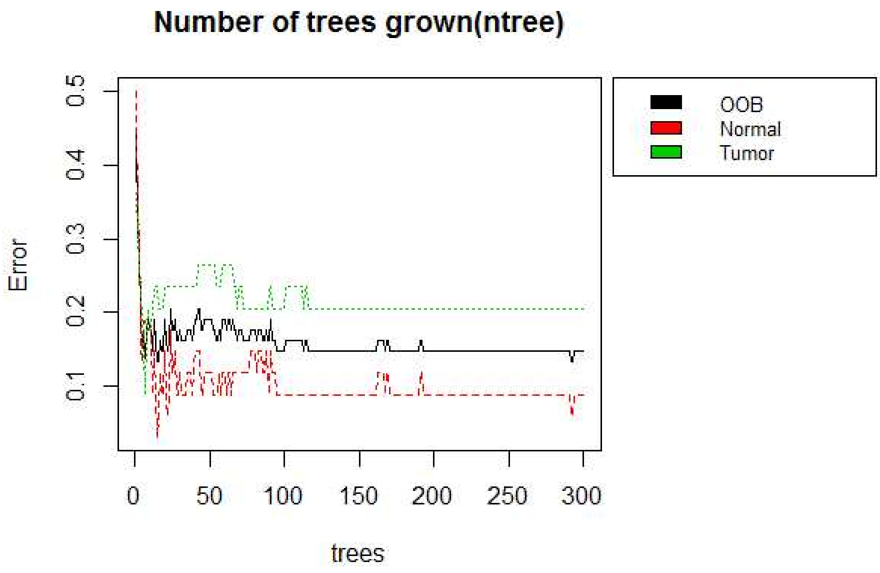


Figure 19. Random Forest OOB ntree

Figure 19 is a graph showing how the error rate of random forest changes with the number of trees using OOB data. As the number of trees increases, the error rate decreases, and it can be seen that the error rate stably converges after about 300 trees. Accordingly, when the random forest model was executed, the number of tree generation is 300 and analysis was performed.

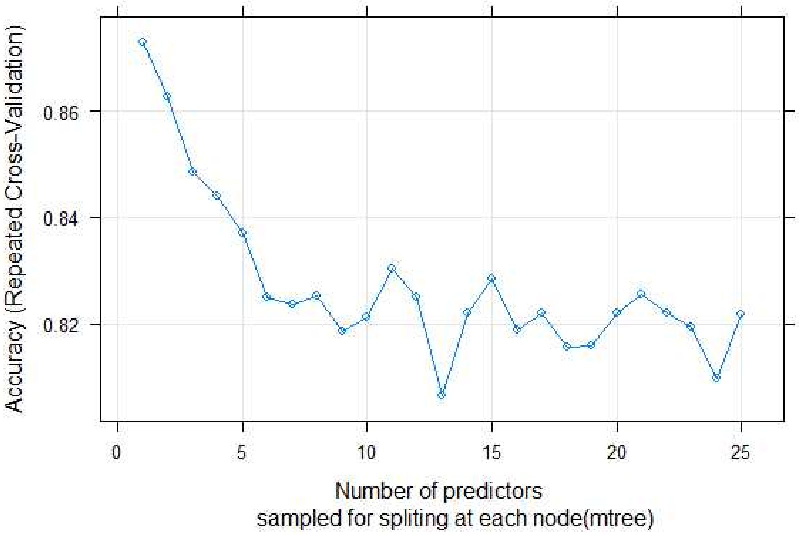
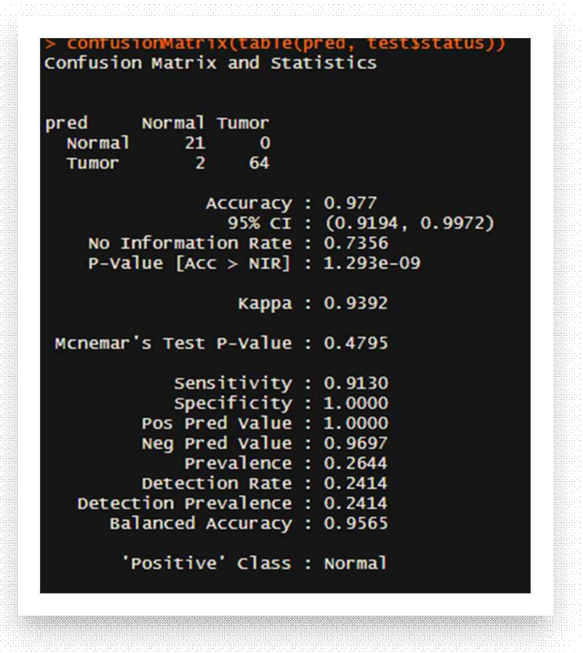


Figure 20. Accuracy for predictors with mtree



```
> confusionMatrix(table(pred, test$status))
Confusion Matrix and Statistics

pred     Normal Tumor
  Normal     21     0
  Tumor       2    64

               Accuracy : 0.977
                 95% CI : (0.9194, 0.9972)
    No Information Rate : 0.7356
    P-Value [Acc > NIR] : 1.293e-09

                  Kappa : 0.9392

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.9130
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9697
             Prevalence : 0.2644
         Detection Rate : 0.2414
   Detection Prevalence : 0.2414
      Balanced Accuracy : 0.9565

       'Positive' Class : Normal
```

Figure 21. Confusion matrix and Statistics

As a result of random forest modeling analysis using TCGA (+SRA) data as a test set and SRA data as a train set by extracting the gene lists extracted through the ROC curve as a target, when the mtree is 1, it had the highest accuracy of 0.977. It was confirmed that it is possible to divide the Normal sample and the Tumor sample at a high rate. Gene Custom analysis and network analysis were performed with selected Gene lists by applying modeling.

5.1.3. Selected Gene Analysis in SRA

Among the genes obtained by performing DEG from TCGA gastric cancer data, Asian-specific genes with AUC 0.8 or higher were extracted, and the gene list obtained by performing DEG with SRA Korean gastric cancer data and common genes applied with random forest modeling were extracted.

Table 4. Common gene list for TCGA and SRA Korean

Gene_Symbol	Gene_Description	Kor_Tumor / Kor_Normal		
		Fold Change	logCPM	bh_pValue
CIP2A	**cell proliferation regulating inhibitor of protein**	**3.500**	**3.233**	**1.9E-18**
LDHD	**lactate dehydrogenase D**	**-3.148**	**5.694**	**9.4E-08**
KIFC1	**kinesin family member C1**	**2.314**	**4.786**	**6.1E-09**
TYMS	thymidylate synthetase	1.893	4.758	6.9E-07
NDRG2	NDRG family member 2	-1.860	6.670	3.2E-06
ME1	malic enzyme 1	-1.683	5.630	6.3E-05
BAIAP2	BAR/IMD domain containing adaptor protein 2	-1.664	6.000	4.5E-04
PVR	PVR cell adhesion molecule	1.444	6.006	1.0E-04
CCDC34	coiled-coil domain containing 34	-1.278	5.086	2.0E-01
AATF	apoptosis antagonizing transcription factor	-1.091	6.237	2.9E-01

In particular, the three genes CIP2A, LDHD, and KIFC1 showed that the direction of regulation was identical in the two datasets (TCGA, SRA), and that the Fold Change was more than twice, and the Adjust p value was also significant. SRA Korean DEG results were marked for 10 genes, and analysis of g;profiler annotation and Cytoscape-based GeneMANIA network is performed.

5.1.4. Differential gene expression according to intensity

Smear Plot. The average expression value of the two groups was high, and it was expressed as a Smear plot to identify 15 extracted genes that are different in the control and test groups. A list of 10 extracted genes was separately displayed.

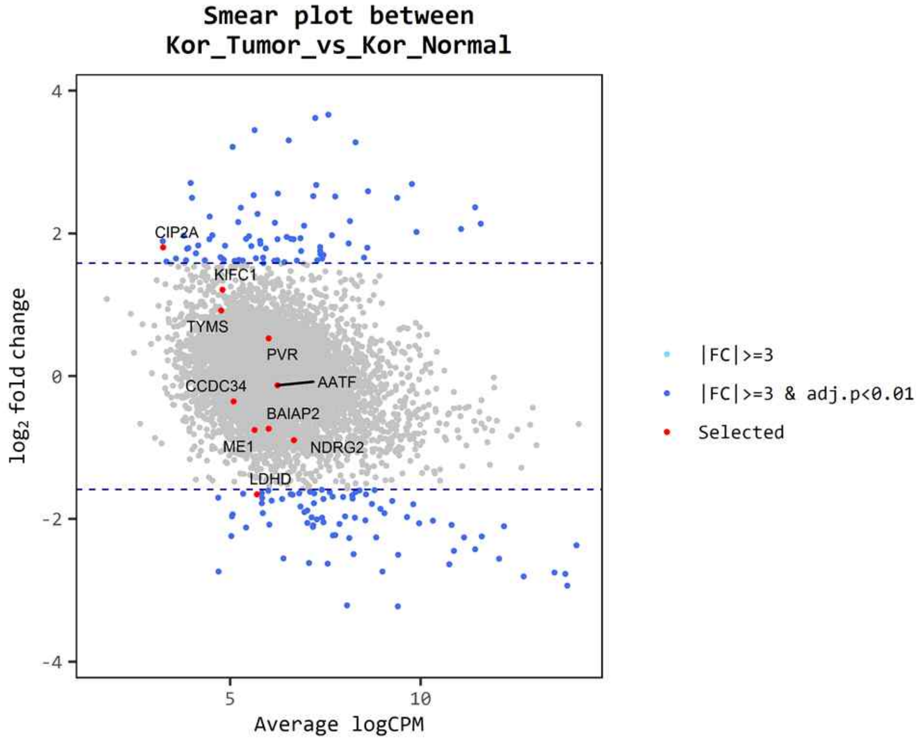


Figure 22. Smear plot with common genes

5.1.5. Volcano plot of expression values between two groups

Figure 23. shows the p-value derived through log2 fold change of expression value between the comparison combinations and the average comparison between the two groups as a volcano plot. Fifteen extracted genes were displayed. A list of 10 extracted genes was separately displayed.

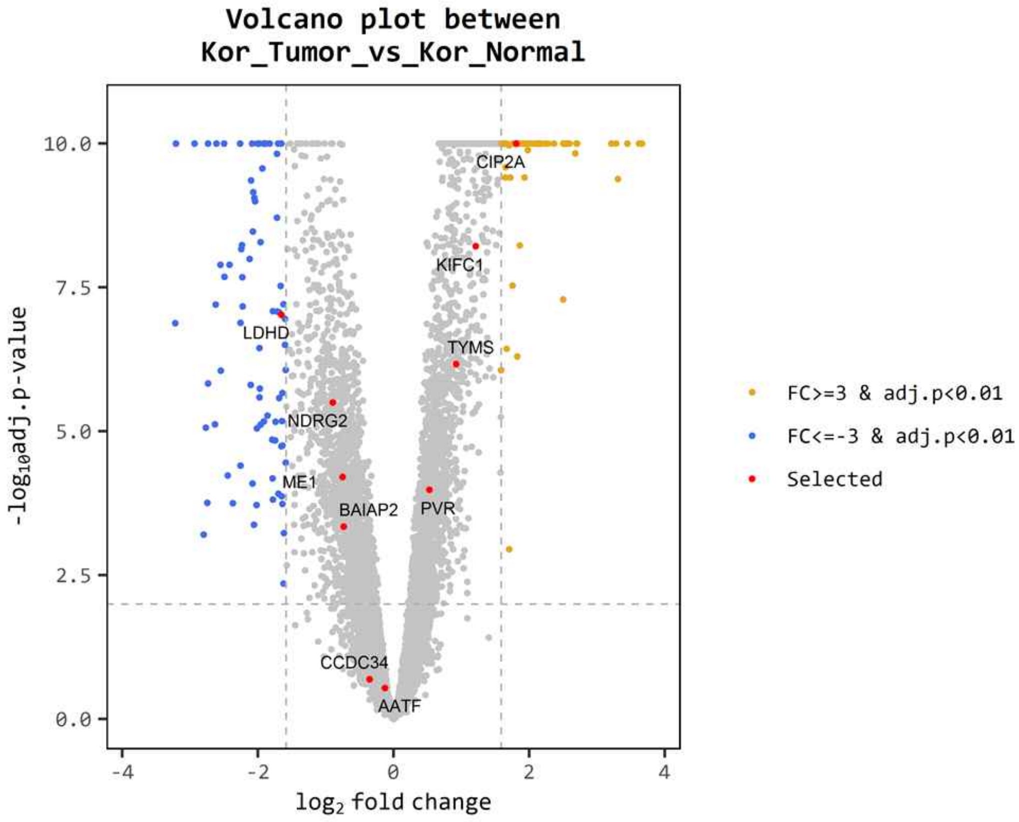


Figure 23. Volcano plot with common genes

5.1.6. Hierarchical Clustering Analysis

For 10 extracted genes, samples and genes with similar expression levels were grouped through hierarchical clustering analysis (Euclidean Distance, Complete Linkage) using the normalized values of each sample. The names of 10 genes were displayed on the right side of the heatmap.

Patterns of Normalized Value according to Normal and Cancer groups can be confirmed for 10 genes.

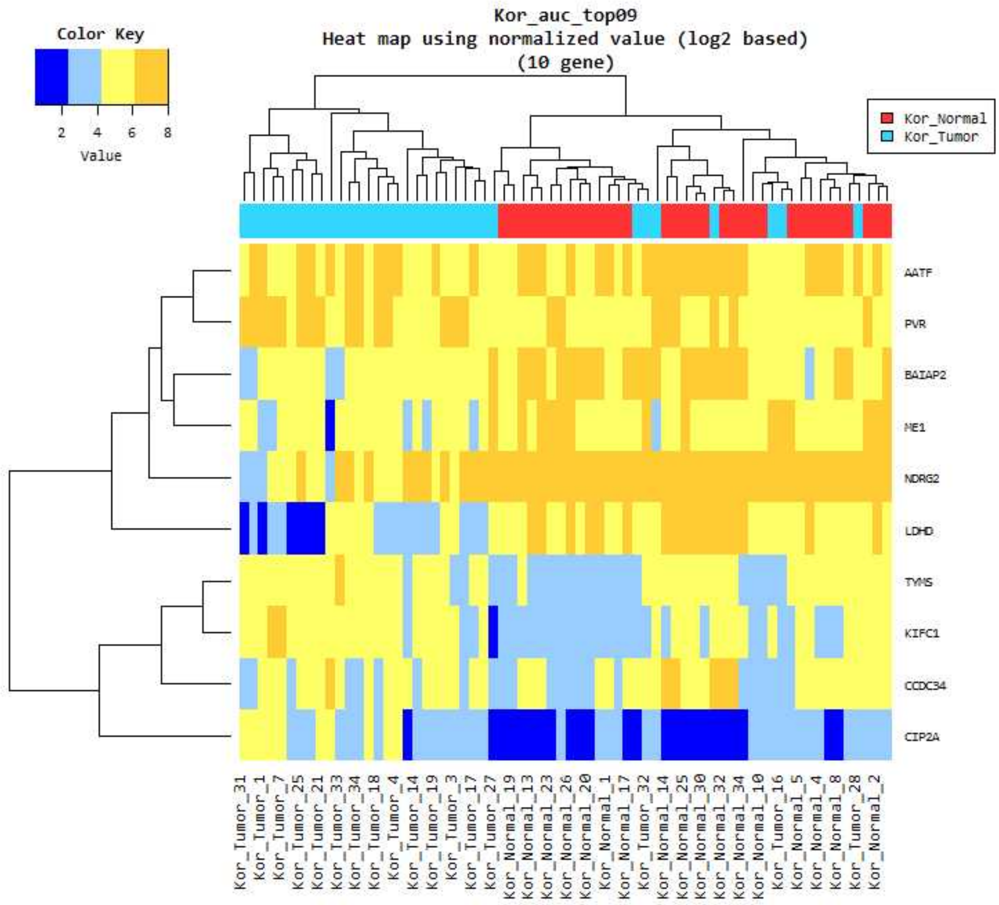


Figure 24. Heatmap with common genes

5.4 Annotation & Network for Korean Specific genes

5.4.1 g;Profiler Annotation

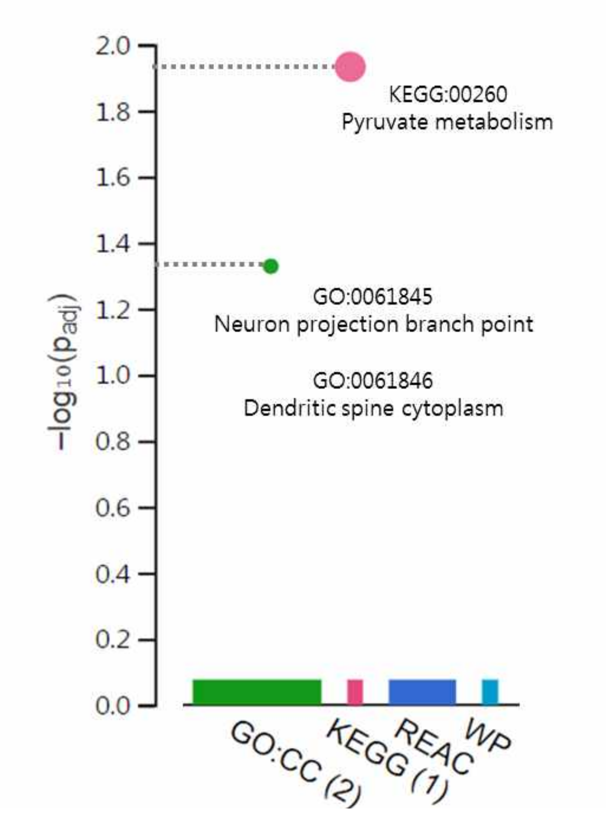


Figure 25. g;Profiler annotation plot

Among g:profiler, g:GOSt is a Gene set enrichment analysis tool that expresses the function by searching the contents of the database with the list of entered genes. The known function of the search gene is mapped to a database to search for a statistically significant function.

The database includes the Ensembl database, Gene Ontology (GO), and KEGG pathway.

In this study, the related functions were confirmed from a total of 3 sources for 10 extracted genes.

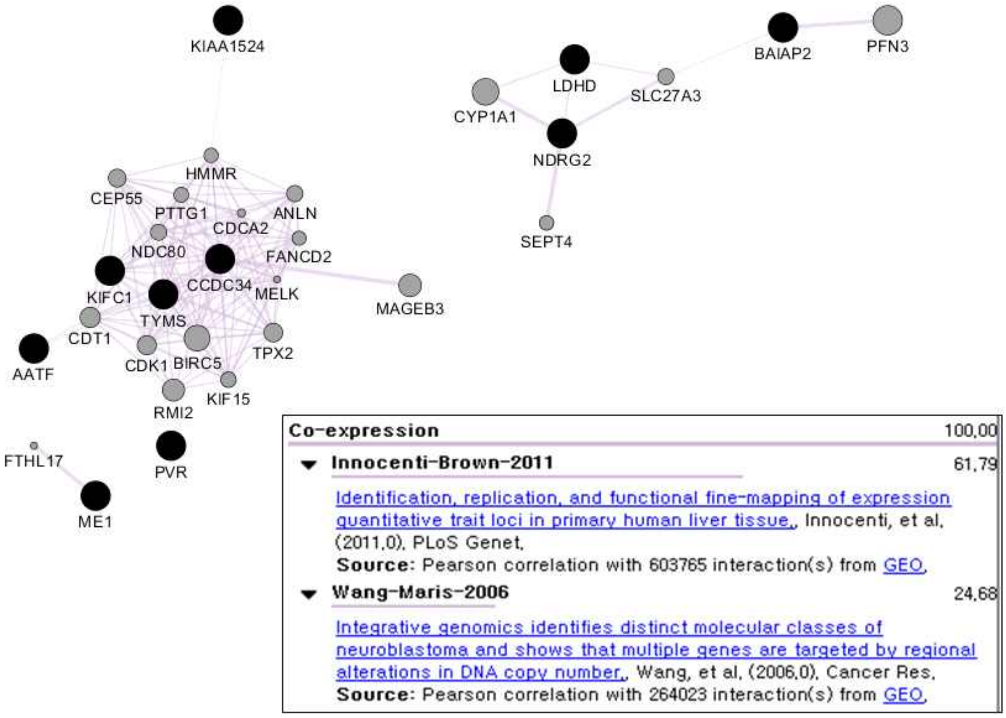
Table 5. g;Profiler DB annotation list

ID	Source	Term ID	Term Name	adj_p
1	KEGG	KEGG:00620	Pyruvate metabolism	1.175×10-2
2	GO:CC	GO:0061846	dendritic spine cytoplasm	4.993×10-2
3	GO:CC	GO:0061845	neuron projection branch point	4.993×10-2

5.4.2 GeneMANIA in Cytoscape

Cytoscape open source software was used as a tool to view the Molecular interaction network of Genes or to view pathways. It is a program that can link various ohmic data such as annotation and expression as well as network and pathway visualization.

Among many plugins of Cytoscape, GeneMANIA network analysis plugin, which is frequently used for Gene-Gene interaction analysis, was used in this study. GeneMANIA is a gene interaction analysis and visualization tool that analyzes 597,392,998 interactions of 163,599 genes in 9 organisms. Genes extracted through RNA-seq DEG analysis are used as inputs to find nearby genes on the network and to find and analyze several networks.



Co-expression 100,00

▼ **Innocenti-Brown-2011** 61,79

Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue., Innocenti, et al, (2011,0), PLoS Genet,
Source: Pearson correlation with 603765 interaction(s) from GEO,

▼ **Wang-Maris-2006** 24,68

Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number., Wang, et al, (2006,0), Cancer Res,
Source: Pearson correlation with 264023 interaction(s) from GEO,

Figure 26. GeneMANIA network plot

Table 6. GeneMANIA GO annotation list

Q-Value	Coverage	GO Annotation
0.072	**3/23**	**Regulation of transcription involved in G1/S transition of mitotic cell cycle**
0.072	**5/180**	**Mitosis**
0.16	4/112	Cytokinesis
0.20	5/256	Nulear division
0.24	5/281	Organelle fission
0.40	2/11	Regulation of exit from mitosis
0.71	3/81	Midbody
0.74	2/18	Mitotic cytokinesis
0.74	2/17	Exit from mitosis

As a GeneMania Network analysis, We can draw a network as shown in Figure 26, and see the mechanism showing the most significant result among the searched networks.

Cytoscape-based GeneMania analysis on the 10 selected genes, it was confirmed that it has the network closest to the two studies.

(1) "Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue.. Innocenti, et al. (2011.0). PLoS Genet."
(2) "Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number.. Wang, et al. (2006.0). Cancer Res."

Also, through Table 5, it was possible to check the Gene Function matching Gene Ontology (GO).

VI. Conclusion and Discussion

Gastric cancer is one of the most prevalent cancers worldwide, and the incidence and mortality rates are increasing, especially according to modern people's eating habits and busy lifestyles. In particular, in the case of Koreans targeted for this study, gastric cancer occupied the number one cause of death among all cancer groups, and the number of cancer deaths was the largest in 2018. It is necessary to discover Korean specific gastric cancer biomarkers to predict the diagnosis of Korean gastric cancer through the results of the study of the American Cancer Control Association (ACSI, 2009) that the incidence and cause of cancer are different for each race. We wanted to see the marker by comparing gene expression differences.

NGS-based RNA-seq analysis is used as a method for viewing gene expression differences at RNA level. In order to discover biomarkers of Korean Gastric Cancer related to RNA expression, analysis is conducted using TCGA Gastric cancer, RNA expression data and NCBI SRA Korean FASTQ data.

In the TCGA data, 359 RNA-seq and Read-Count data of three races (Asian, White, and Black) from the Gastric Cancer data are received and proceeded from DEG analysis. Analysis is conducted with 23 Normal samples and 336 Tumor Samples by race, and differential expression genes are extracted for Asian races. The limitation of this study is that the analysis is conducted with the Normal group, which combined the three races due to the small number of Normal samples of Asian and Black races.

SRA data received FASTQ data from the Korean gastric cancer project (Accession: PRJNA435914) and RNA-seq analysis is conducted through the server itself. The RNA-seq pipeline is analyzed in the same way as the pipeline of TCGA RNA-seq, and DEG analysis is also conducted by grouping 34 samples of each of Normal and Cancer samples.

Twenty-eight Korean-specific biomarkers of gastric cancer genes based on TCGA gastric cancer data are identified by substituting a list of 28 differential expression genes of Asian races selected as AUC filters into the list of Korean differential expression genes to which a random forest model is applied. Genes showing the most statistically significant values are three genes, CIP2A, LDHD, and KIFC1, respectively. Annotation of cell proliferation regulation inhibitor of protein, lactate dehydrogenase D, and Kinesin family member C1 is confirmed. The CIP2A and KIFC1 genes are up-regulated DEGs in the Gastric Cancer patient group than the general population. LDHD gene can be confirmed that gastric cancer patient group had lower-regulated DEGs than the general population. These three genes CIP2A, KIFC1, and LDHD can be used as biomarkers for Korean gastric cancer-specific genes.

In addition, by searching g;profiler Annotation and GeneMANIA Network for the 10 genes discovered, functions related to pathways containing Korean-specific genes are derived. Through the g;Profiler Annotation, a total of three processes are identified. The Pyruvate metabolism pathway of KEGG:00260 ID, GO:0061846 dendritic spine cytoplasm, and GO:0061845 neuron projection branch points showed significant adjust p-value.

GeneMANIA Network analysis confirmed a significant result in a total of two co-expressions. The corresponding co-expression is “Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue.. Innocenti, et al. (2011.0). PLoS Genet.” Research and “Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. Wang, et al. (2006.0). Cancer Res.” It is confirmed that the two studies are closely related.

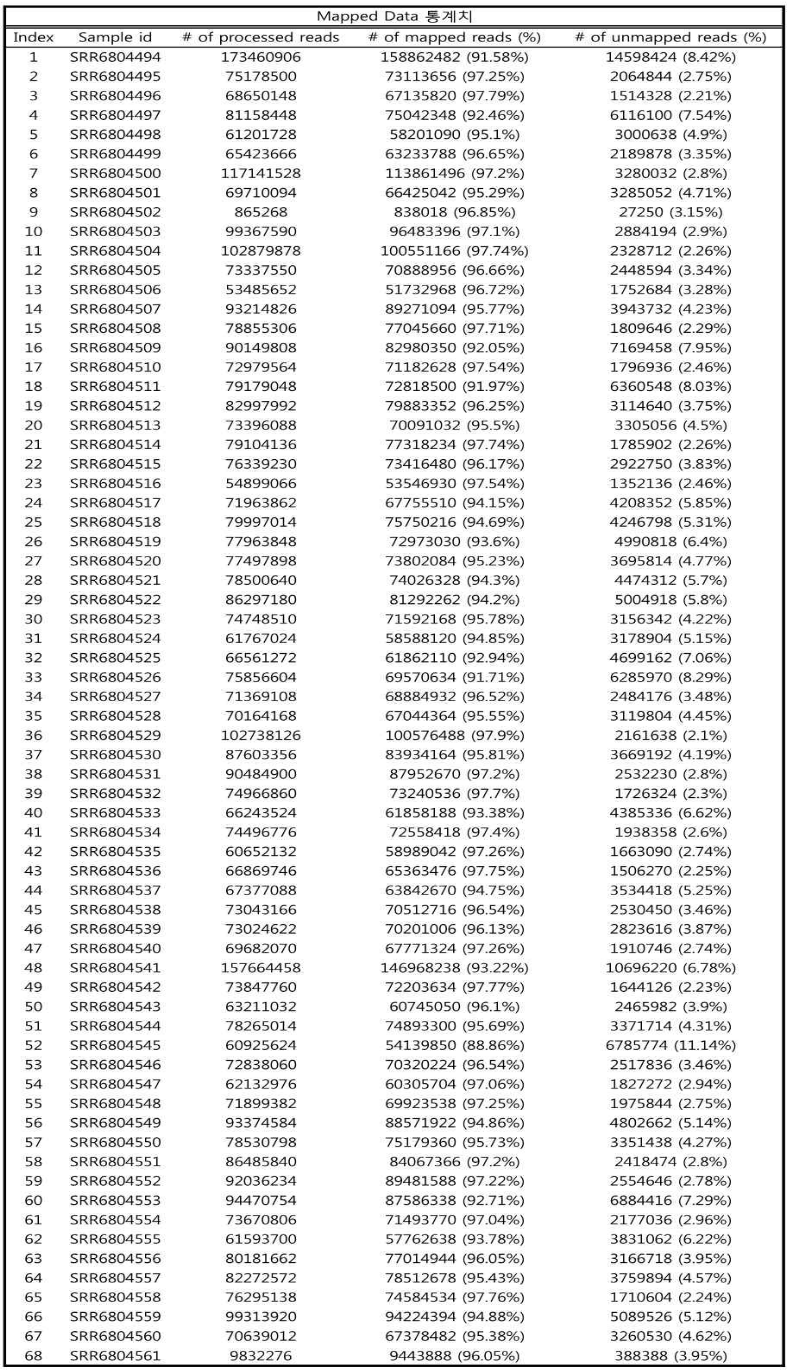
Three Korean-specific genes (CIP2A, LDHD, KIPFC1) and three g;profiler results (KEGG:00260, GO:0061846, GO:0061845), and two GeneMANIA Network results extracted.

In particular, gastric cancer appears to be more common in Asians than in whites and blacks. Through this study It is expected to be used as a biomarker for predicting stomach cancer diagnosis in Koreans.

Table 7. Data stat with RawData and Trimmed Data

Data Preprocessing													
RawData 데이터 기초 통계치							Trimmed Data 통계치						
Index	Sample id	Total read bases*	Total reads	GC(%)	Q20(%)	Q30(%)	Index	Sample id	Total read bases*	Total reads	GC(%)	Q20(%)	Q30(%)
1	SRR6804494	27349776850	181124350	52.17	96.74	92.85	1	SRR6804494_trim	24106275559	173460906	52.37	98.7	95.72
2	SRR6804495	12402893870	82138370	49.87	93.78	88.12	2	SRR6804495_trim	10330843994	75178500	50.43	97.99	93.83
3	SRR6804496	10869964520	71986520	48.88	95.71	90.59	3	SRR6804496_trim	9801556159	68650148	49	98.14	94.1
4	SRR6804497	12752956868	84456668	51.62	96.48	91.98	4	SRR6804497_trim	11710091149	81158448	51.85	98.42	94.82
5	SRR6804498	10262316360	67962360	50.13	92.01	84.55	5	SRR6804498_trim	8405697287	61201728	50.69	97.1	91.36
6	SRR6804499	10403872216	68899816	49.2	96.02	91.39	6	SRR6804499_trim	9257660019	65423666	49.23	98.36	94.78
7	SRR6804500	18364512790	121619290	50.39	96.81	92.83	7	SRR6804500_trim	16591526722	117141528	50.55	98.65	95.54
8	SRR6804501	11160012266	73907366	51.22	95.68	90.81	8	SRR6804501_trim	9900882659	69710094	51.53	98.36	94.66
9	SRR6804502	138470020	917020	50.92	96.19	91.48	9	SRR6804502_trim	122585365	865268	51.06	98.41	94.78
10	SRR6804503	15802271404	104650804	51.55	96.46	92.08	10	SRR6804503_trim	13967760629	99367590	51.71	98.57	95.2
11	SRR6804504	16284195152	107842352	49.36	96.27	91.67	11	SRR6804504_trim	14794023428	102879878	49.46	98.39	94.85
12	SRR6804505	11621157206	76961306	50.69	96.37	92.04	12	SRR6804505_trim	10338596235	73337550	50.85	98.53	95.22
13	SRR6804506	8600311304	56955704	50.06	95.33	90.02	13	SRR6804506_trim	7512914869	53485652	50.22	98.28	94.27
14	SRR6804507	14954918898	99039198	49.95	95.54	90.24	14	SRR6804507_trim	13229420068	93214826	50.07	98.26	94.26
15	SRR6804508	12853505956	85122556	50.37	95.96	91.54	15	SRR6804508_trim	11218515869	78855306	50.51	98.54	95.22
16	SRR6804509	14605443894	96724794	53.4	95.84	91.17	16	SRR6804509_trim	12882691923	90149808	53.77	98.45	94.89
17	SRR6804510	11470155998	75961298	51.01	96.15	91.38	17	SRR6804510_trim	10205600068	72979564	51.1	98.3	94.55
18	SRR6804511	13434519226	88970326	51.98	94.5	89.03	18	SRR6804511_trim	11146136835	79179048	52.22	98.17	94.17
19	SRR6804512	13019964128	86224928	50.82	96.93	93.13	19	SRR6804512_trim	11644981869	82997992	50.98	98.73	95.78
20	SRR6804513	11579122732	76682932	50.18	96.62	92.67	20	SRR6804513_trim	10258028368	73396088	50.37	98.68	95.64
21	SRR6804514	13076281692	86597892	50.72	95.58	91.09	21	SRR6804514_trim	11125391258	79104136	50.94	98.55	95.29
22	SRR6804515	12291802264	81402664	51.78	96	91.39	22	SRR6804515_trim	10868677974	76339230	52.02	98.45	94.97
23	SRR6804516	8665907818	57390118	49.97	96.1	91.19	23	SRR6804516_trim	7772035138	54899066	50.06	98.35	94.54
24	SRR6804517	11533672034	76381934	51.33	95.3	89.55	24	SRR6804517_trim	10232455629	71963862	51.56	97.95	93.47
25	SRR6804518	12741207256	84378856	52.57	95.86	91.35	25	SRR6804518_trim	11039385498	79997014	53.02	98.5	95.18
26	SRR6804519	12803626126	84792226	52.42	95.14	90.65	26	SRR6804519_trim	10704592881	77963848	53	98.58	95.42
27	SRR6804520	12134954638	80363938	50.28	96.8	92.95	27	SRR6804520_trim	10643013897	77497898	50.38	98.71	95.83
28	SRR6804521	12442683578	82401878	50.34	96.22	92.1	28	SRR6804521_trim	10808666422	78500640	50.59	98.63	95.59
29	SRR6804522	13506288318	89445618	50.06	96.73	92.79	29	SRR6804522_trim	11929771187	86297180	50.11	98.67	95.71
30	SRR6804523	11695163212	77451412	50.35	96.86	93.07	30	SRR6804523_trim	10291024696	74748510	50.44	98.74	95.9
31	SRR6804524	9951018384	65900784	51	96	91.67	31	SRR6804524_trim	8531054373	61767024	51.18	98.56	95.37
32	SRR6804525	10813202714	71610614	52.56	95.61	91.19	32	SRR6804525_trim	9114911648	66561272	53.08	98.58	95.4
33	SRR6804526	12099153746	80126846	51.23	95.89	91.37	33	SRR6804526_trim	10524783206	75856604	51.5	98.45	95.06
34	SRR6804527	11202213142	74186842	50.38	96.68	92.77	34	SRR6804527_trim	9784967193	71369108	50.5	98.7	95.81
35	SRR6804528	11202300118	74187418	51.62	95.66	90.91	35	SRR6804528_trim	9607940657	70164168	51.81	98.41	94.94
36	SRR6804529	16300649018	107951318	50.33	95.51	89.94	36	SRR6804529_trim	14659882900	102738126	50.35	97.99	93.58
37	SRR6804530	14226453826	94214926	49.93	94.78	88.96	37	SRR6804530_trim	12415592014	87603356	50.03	97.93	93.43
38	SRR6804531	14212719470	94123970	51.14	96.55	92.17	38	SRR6804531_trim	12964119660	90484900	51.3	98.45	94.94
39	SRR6804532	11832772532	78362732	51.38	96.18	91.52	39	SRR6804532_trim	10678161782	74966860	51.63	98.36	94.69
40	SRR6804533	10477154932	69385132	51.03	96.53	92.31	40	SRR6804533_trim	9330969190	66243524	51.2	98.56	95.31
41	SRR6804534	11802656186	78163286	50.03	96.37	91.95	41	SRR6804534_trim	10625466854	74496776	50.18	98.48	95.09
42	SRR6804535	9662412386	63989486	51.02	96.38	91.88	42	SRR6804535_trim	8721681408	60652132	51.15	98.43	94.87
43	SRR6804536	10568207328	69988128	50.99	94.72	88.43	43	SRR6804536_trim	9115680486	66869746	51.1	97.58	92.62
44	SRR6804537	10627894910	70383410	50.6	96.27	91.66	44	SRR6804537_trim	9647060921	67377088	50.79	98.37	94.7
45	SRR6804538	11476480180	76003180	50.86	96.55	92.21	45	SRR6804538_trim	10459627206	73043166	51.04	98.47	95
46	SRR6804539	11557602514	76540414	51.74	96.54	92.62	46	SRR6804539_trim	10149115060	73024622	51.97	98.71	95.75
47	SRR6804540	11101989308	73523108	50.57	95.99	91.54	47	SRR6804540_trim	9636551815	69682070	50.72	98.49	95.19
48	SRR6804541	26271709028	173984828	52.3	94.89	89.73	48	SRR6804541_trim	22583001363	157664458	52.75	98.37	94.54
49	SRR6804542	12294313394	81419294	48.56	94.09	88.28	49	SRR6804542_trim	10500341618	73847760	49.01	98.02	93.66
50	SRR6804543	10417191624	68988024	51	95.41	90.94	50	SRR6804543_trim	8718606398	63211032	51.34	98.57	95.41
51	SRR6804544	12357245362	81836062	52.3	95.36	89.82	51	SRR6804544_trim	10464654572	78265014	52.46	98.03	93.87
52	SRR6804545	10666292700	70637700	50.55	92.57	87.28	52	SRR6804545_trim	8444719931	60925624	51.8	98.35	94.79
53	SRR6804546	11426680984	75673384	50.48	96.76	92.92	53	SRR6804546_trim	9934905601	72838060	50.58	98.74	95.89
54	SRR6804547	9766500914	64678814	50.41	96.66	92.69	54	SRR6804547_trim	8735397106	62132976	50.56	98.65	95.56
55	SRR6804548	11290804540	74773540	49.69	96.74	92.74	55	SRR6804548_trim	10130293352	71899382	49.81	98.63	95.52
56	SRR6804549	15123557812	100156012	52.25	95.59	91.38	56	SRR6804549_trim	12738747332	93374584	52.85	98.66	95.66
57	SRR6804550	12631709002	83653702	51.24	95.87	91.6	57	SRR6804550_trim	10756013427	78530798	51.58	98.6	95.49
58	SRR6804551	13605365156	90101756	52.17	96.69	92.74	58	SRR6804551_trim	12019529990	86485840	52.46	98.69	95.66
59	SRR6804552	14361127404	95106804	49.72	97.01	93.27	59	SRR6804552_trim	12683089055	92036234	49.7	98.72	95.9
60	SRR6804553	15063213682	99756382	51.8	96.2	92.05	60	SRR6804553_trim	13009669488	94470754	52.23	98.62	95.57
61	SRR6804554	11664275256	77246856	48.94	96.19	91.95	61	SRR6804554_trim	10194994221	73670806	49.07	98.53	95.35
62	SRR6804555	9879821280	65429280	51.56	93.82	87.38	62	SRR6804555_trim	8161689560	61593700	51.98	97.61	92.9
63	SRR6804556	12624001962	83602662	51.42	96.46	92.34	63	SRR6804556_trim	10967210325	80181662	51.62	98.62	95.56
64	SRR6804557	13370014442	88543142	52.78	95.48	90.94	64	SRR6804557_trim	11709629920	82272572	53.41	98.52	95.13
65	SRR6804558	12066089880	79907880	50.06	95.33	89.65	65	SRR6804558_trim	10830938865	76295138	50.09	97.95	93.47
66	SRR6804559	15707598934	104023834	51.17	95.68	90.16	66	SRR6804559_trim	14088176387	99313920	51.37	97.99	93.58
67	SRR6804560	11072517128	73327928	54.71	96.9	92.77	67	SRR6804560_trim	10211184298	70639012	54.95	98.61	95.32
68	SRR6804561	1580487404	10466804	52.41	96.46	92.78	68	SRR6804561_trim	1417770609	9832276	52.76	98.87	96.12

Table 8. Data stat with mapped reads



Mapped Data 통계치				
Index	Sample id	# of processed reads	# of mapped reads (%)	# of unmapped reads (%)
1	SRR6804494	173460906	158862482 (91.58%)	14598424 (8.42%)
2	SRR6804495	75178500	73113656 (97.25%)	2064844 (2.75%)
3	SRR6804496	68650148	67135820 (97.79%)	1514328 (2.21%)
4	SRR6804497	81158448	75042348 (92.46%)	6116100 (7.54%)
5	SRR6804498	61201728	58201090 (95.1%)	3000638 (4.9%)
6	SRR6804499	65423666	63233788 (96.65%)	2189878 (3.35%)
7	SRR6804500	117141528	113861496 (97.2%)	3280032 (2.8%)
8	SRR6804501	69710094	66425042 (95.29%)	3285052 (4.71%)
9	SRR6804502	865268	838018 (96.85%)	27250 (3.15%)
10	SRR6804503	99367590	96483396 (97.1%)	2884194 (2.9%)
11	SRR6804504	102879878	100551166 (97.74%)	2328712 (2.26%)
12	SRR6804505	73337550	70888956 (96.66%)	2448594 (3.34%)
13	SRR6804506	53485652	51732968 (96.72%)	1752684 (3.28%)
14	SRR6804507	93214826	89271094 (95.77%)	3943732 (4.23%)
15	SRR6804508	78855306	77045660 (97.71%)	1809646 (2.29%)
16	SRR6804509	90149808	82980350 (92.05%)	7169458 (7.95%)
17	SRR6804510	72979564	71182628 (97.54%)	1796936 (2.46%)
18	SRR6804511	79179048	72818500 (91.97%)	6360548 (8.03%)
19	SRR6804512	82997992	79883352 (96.25%)	3114640 (3.75%)
20	SRR6804513	73396088	70091032 (95.5%)	3305056 (4.5%)
21	SRR6804514	79104136	77318234 (97.74%)	1785902 (2.26%)
22	SRR6804515	76339230	73416480 (96.17%)	2922750 (3.83%)
23	SRR6804516	54899066	53546930 (97.54%)	1352136 (2.46%)
24	SRR6804517	71963862	67755510 (94.15%)	4208352 (5.85%)
25	SRR6804518	79997014	75750216 (94.69%)	4246798 (5.31%)
26	SRR6804519	77963848	72973030 (93.6%)	4990818 (6.4%)
27	SRR6804520	77497898	73802084 (95.23%)	3695814 (4.77%)
28	SRR6804521	78500640	74026328 (94.3%)	4474312 (5.7%)
29	SRR6804522	86297180	81292262 (94.2%)	5004918 (5.8%)
30	SRR6804523	74748510	71592168 (95.78%)	3156342 (4.22%)
31	SRR6804524	61767024	58588120 (94.85%)	3178904 (5.15%)
32	SRR6804525	66561272	61862110 (92.94%)	4699162 (7.06%)
33	SRR6804526	75856604	69570634 (91.71%)	6285970 (8.29%)
34	SRR6804527	71369108	68884932 (96.52%)	2484176 (3.48%)
35	SRR6804528	70164168	67044364 (95.55%)	3119804 (4.45%)
36	SRR6804529	102738126	100576488 (97.9%)	2161638 (2.1%)
37	SRR6804530	87603356	83934164 (95.81%)	3669192 (4.19%)
38	SRR6804531	90484900	87952670 (97.2%)	2532230 (2.8%)
39	SRR6804532	74966860	73240536 (97.7%)	1726324 (2.3%)
40	SRR6804533	66243524	61858188 (93.38%)	4385336 (6.62%)
41	SRR6804534	74496776	72558418 (97.4%)	1938358 (2.6%)
42	SRR6804535	60652132	58989042 (97.26%)	1663090 (2.74%)
43	SRR6804536	66869746	65363476 (97.75%)	1506270 (2.25%)
44	SRR6804537	67377088	63842670 (94.75%)	3534418 (5.25%)
45	SRR6804538	73043166	70512716 (96.54%)	2530450 (3.46%)
46	SRR6804539	73024622	70201006 (96.13%)	2823616 (3.87%)
47	SRR6804540	69682070	67771324 (97.26%)	1910746 (2.74%)
48	SRR6804541	157664458	146968238 (93.22%)	10696220 (6.78%)
49	SRR6804542	73847760	72203634 (97.77%)	1644126 (2.23%)
50	SRR6804543	63211032	60745050 (96.1%)	2465982 (3.9%)
51	SRR6804544	78265014	74893300 (95.69%)	3371714 (4.31%)
52	SRR6804545	60925624	54139850 (88.86%)	6785774 (11.14%)
53	SRR6804546	72838060	70320224 (96.54%)	2517836 (3.46%)
54	SRR6804547	62132976	60305704 (97.06%)	1827272 (2.94%)
55	SRR6804548	71899382	69923538 (97.25%)	1975844 (2.75%)
56	SRR6804549	93374584	88571922 (94.86%)	4802662 (5.14%)
57	SRR6804550	78530798	75179360 (95.73%)	3351438 (4.27%)
58	SRR6804551	86485840	84067366 (97.2%)	2418474 (2.8%)
59	SRR6804552	92036234	89481588 (97.22%)	2554646 (2.78%)
60	SRR6804553	94470754	87586338 (92.71%)	6884416 (7.29%)
61	SRR6804554	73670806	71493770 (97.04%)	2177036 (2.96%)
62	SRR6804555	61593700	57762638 (93.78%)	3831062 (6.22%)
63	SRR6804556	80181662	77014944 (96.05%)	3166718 (3.95%)
64	SRR6804557	82272572	78512678 (95.43%)	3759894 (4.57%)
65	SRR6804558	76295138	74584534 (97.76%)	1710604 (2.24%)
66	SRR6804559	99313920	94224394 (94.88%)	5089526 (5.12%)
67	SRR6804560	70639012	67378482 (95.38%)	3260530 (4.62%)
68	SRR6804561	9832276	9443888 (96.05%)	388388 (3.95%)

Reference

[1] 국가통계포털, 사망원인통계(국가승인통계 제101054호), 통계청. 2017

[2] 국가통계포털, 암등록통계(국가승인통계 117044호), 보건복지부. 2017

[3] 중앙암등록본부, 국가암등록통계 참고 자료, 보건복지부. 2015

[4] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 2013, 29.1: 15-21.

[5] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, et al. The sequence alignment/map format and SAMtools. Bioinformatics, 2009, 25.16:2078-2079.

[6] BOLGER, Anthony M.; LOHSE, Marc; USADEL, Bjoern. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 2014.

[7] ANDERS, Simon; PYL, Paul Theodor; HUBER, Wolfgang. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics, 2014, btu638.

[8] DEPRISTO, Mark A., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics, 2011, 43.5: 491-498.

[9] MCKENNA, Aaron, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 2010, 20.9: 1297-1303.

[10] Yunshun Chen, Davis McCarthy, Matthew Ritchie, et al. edgeR: differential expression analysis of digital gene expression data, Bioconducor, 2019

[11] Robinson, MD, and Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25, 2010.

[12] Hajian-Tilald K, “Reveiver Operating Characteristic(ROC) curve Analysis for Medical Diagnostic Test Evaluation”, Caspian, J Intern Med, 4(2), 627-635. 2013.

[13] 유진은, “랜덤 포레스트: 의사결정나무의 대안으로서의 데이터 마이닝 기법”, Journal of Educational Evaluation, 28(2):427-448.99a, 2015

[14] Yoo, Jin Soo, 2019, “Diagnostic Factor and Prediction Model for Stomach Cancer Using TCGA Database”, Chungbuk National University.

[15] Kim, Jong Bum, 2014, “Identification of Hypermethylated genes and their impact on Colorectal Cancer”, Graduate School of Soongsil University

[16] Tark, Yeon Jeong, 2014, “Understanding of Difference between Korean lung cancer population and others using NGS exome sequencing”, Graduate School of Soongsil University

[17] Kijin Yu, 2019, “Pan-Cancer Classification on gene Expression Data by Maching Learning”, Chungbuk National University.

[18] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, et al, April 5, 2018, “An Integrated TCGA Pna-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics”, Cell 173, 400-416

[19] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, et al, April 14, 2012, “Discovery of multi-dimensional modules by integrative analysis of cancer genomic data”, Nuclleic Acids Research, Vol. 40, No 19, 9379-9391

[20] Gregory P. Way, Francisco Sanchez-vega, et al, April 3, 2018, “Machine Lrearning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas”, Cell Reports 23, 172-180

[21] Wenjuan Li, Zheng Ge, et al, June 2008, “CIP2A Is Overexpressed in Gastric Cancer and Its Depletion Leads to Impaired Clonogenicity, Senescence, or Differentiation of Tumor Cells”, Clinical Cancer Research, 10.1158/1078-0432.CCR-07-4137

[22] Zhaoxing Li, Zhao Liu, et al, May 25, 2020, “Identifying multiple collagen gene family members as potential gastric cancer biomarkers using integrated bioinformatics analysis”, PeerJ, 10.7717/peerj.9123

[23] Naohide Oue Shoichiro Mukai, et al, April 28, 2016, “Induction of KIFC1 expression in gastric cancer spheroids,”, Oncology Reports, 349-355

국문요약

위암은 전 세계적으로 발병률이 높은 암 중 하나로 특히 현대인들의 식습관과 바쁜 생활습관에 따라 발생률과 사망률이 증가하고 있다. 특히 본 연구에서 대상으로 정한 한국인의 경우, 전체 암 군 중 사망원인의 1위를 위암이 차지하고 있으며, (2018. 사망원인통계) 암 사망자 수도 위암에서 최대 기록을 보이고 있다. 인종 별 암 발병률 및 원인이 다르다는 미 암 통제협회(ACSI, 2009)의 연구 결과를 통해 한국인 위암 진단 예측을 위한 한국인 특정 위암 바이오마커 발굴의 필요하며, 이번 연구에서는 RNA에서 정상군과 위암군의 유전자 발현차이를 비교하여 바이오마커를 보고자 하였다.

RNA 에서의 유전자 발현차이를 보기위한 방법으로 NGS 기반의 RNA-seq 분석법을 활용하였다. RNA 발현 관련해서 한국인 위암 바이오마커를 발굴하기 위해 TCGA 데이터베이스의 RNA 발현 데이터와 NCBI SRA 데이터의 한국인 FASTQ 데이터를 활용하여 분석을 진행하였다.

TCGA 데이터에서는 위암 데이터 중 세 인종(아시안, 백인, 흑인)의 RNA-seq 발현 자료로 분석하였다. 정상 샘플과 인종별 위암샘플로 분석을 진행하였고 아시안 차별 발현 유전자를 추출하였다. SRA 데이터는 한국인 위암 프로젝트 (Accession: PRJNA435914)의 FASTQ 데이터를 받아 RNA-seq 분석을 진행하였다. RNA-seq 파이프라인은 TCGA RNA-seq의 파이프라인과 동일하게 맞춰 분석을 진행하였고, 정상 샘플과 위암 샘플 각각 그룹으로 하여 분석하였다.

AUC 필터로 추려진 아시안 인종의 차별 발현 유전자를 대상으로 랜덤포레스트 모델이 적용된 한국인 차별 발현 유전자 리스트에 대입하여 TCGA 위암 데이터를 기반으로 한 한국인 특화 위암 유전자 바이오마커를 발굴하였다. 통계적으로 가장 유의한 값을 보이는 유전자는 CIP2A, LDHD, KIPFC1 세 개의 유전자로 확인할 수 있었고 . 또한 발굴된 10개의 유전자에 대해 g;profiler Annnotation과 GeneMANIA 네트워크 검색을 하여 한국인 특화 유전자가 포함된 pathway에 관한 기능들을 도출하였다.

본 연구를 통해 추출된 TCGA 기반의 한국인 특화 유전자 3개 (CIP2A, LDHD, KIPFC1)와 3개의 g;profiler 결과 (KEGG:00260, GO:0061846, GO:0061845), 2개의 GeneMANIA Network 결과가 향후 한국인을 대상으로 하는 위암 진단 예측을 위한 바이오마커로 활용될 수 있기를 기대해보며, 이러한 RNA-seq 연구를 통해 위암 환자에 있어 인종별로의 차이를 인지하여 보다 안전하고 정확한 바이오마커가 점점 늘어나길 기대해본다.