Survival rate and chronic diseases of TCGA cancer and KoGES normal samples by clustering for DNA methylation

Jeong-An Gim (✉ vitastar@korea.ac.kr)
Korea University Medical Center

Research Article

Keywords: TCGA, KoGES, DNA methylation, Survival rate, Complex diseases

Posted Date: April 19th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2826409/v1

License: ☕️️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Insights from public DNA methylation data derived from cancer or normal people can be obtained by machine learning. The goal is to determine the methylation pattern for predicting the prognosis for cancer patients and correcting lifestyles for normal people.

Methods

DNA methylation data were obtained from 446 healthy participants of 11 TCGA carcinomas, the Korean Genome Epidemiology Study (KoGES). For males and females, respectively, data from TCGA and KoGES were merged. To correct for the batch effect, R's ComBat function is used. Using the K-mean clustering (k = 3), the survival rates and chronic diseases were confirmed between the three groups of TCGA and KoGES.

Results

In a total of 82 female pancreatic cancer (TCGA-PAAD) patients and 220 normal participants, cluster 1 and 3 groups were compared. The survival rate in cluster 3 was significantly lower than in cluster 1. Among normal participants, the odds ratio according to the presence or absence of hypertension in two clusters was 2.18.

Conclusion

Based on the public DNA methylation and clinical data of normal participants and cancer patients, an analysis pipeline that integrates and clusters were provided. As a result of clustering, genes that explain survival and chronic disease were presented.

Introduction

The cancer genome atlas (TCGA) is a cancer genome database and provides information on variants, copy number variations (CNVs), RNA-sequencing (RNA-seq), small RNA-seq (sRNA-seq), and DNA methylation for a total of 33 carcinomas. To understand the genetic patterns that cause cancer, TCGA aims to provide clues that can diagnose, treat, and prevent cancer [1, 2]. TCGA was a pilot project for three years in 2006, and omics data were analyzed in three carcinomas (glioblastoma, lung cancer, and ovarian cancer), and omics data for a total of 33 carcinomas, including 10 rare cancers, were analyzed from 2009 to 2014. The data has been completed and open to the public. By secondary analysis of the published data, it was possible to have a deep understanding according to the subtype and histological patterns of cancer [1, 3, 4].

Korean Genome and Epidemiology Study (KoGES) is a cohort study to find risk factors for chronic diseases in Koreans. Among them, the Ansung-Ansan cohort was started in 2001 for the general
population aged 40–60, and follow-up was conducted every 2 years after the base survey of 10,030 people [5]. Participants filled out a questionnaire to check their current health status and lifestyle, and received a health check-up. Blood 30ml and urine 15ml were collected and used for clinical examination. A genotype analysis was performed on a total of 8,840 people in the baseline survey [6, 7], and metabolite analysis was performed on 2,580 people in the second follow-up after 4 years of the baseline survey [8]. DNA methylation analysis was performed on 446 people in the baseline survey, and additional analysis was performed on 50 out of 446 people in the fourth follow-up after 8 years to observe changes in DNA methylation patterns according to time and lifestyle [9, 10].

DNA methylation is a biological process in which a methyl group is added to the 5th carbon of cytosine constituting DNA. It is an epigenetic mechanism that induces biological changes without sequence change [11]. DNA methylation generally plays a role in suppressing gene expression and is known to be involved in development, X chromosome imprinting, aging, and carcinogenesis. DNA methylation varies with sex, age, disease state, and lifestyle [12, 13]. By dietary intervention or DNA methylation analysis before and after clinical treatment, a time-dependent change of DNA methylation according to lifestyle was reported [14, 15]. The degree of DNA methylation is presented as a beta value, which is a value between 0 and 1, and the closer to 1, the more CpG methylation occurs. DNA methylation patterns differ according to sex, because women have two X chromosomes, so genomic imprinting occurs in one of them. This shows higher methylation levels in females, which is a reason to analyze females and males separately [9, 15].

DNA methylation patterns are modified by aging, and epigenetic clocks were revealed from the normal participants as well as the KoGES sample, the general population of 40–60 years of age [15–17]. In general, cancer often develops after the age of 50, so this study aims to find clustering patterns of DNA methylation in normal and cancer subjects. For clustering on the degree of DNA methylation, cancer patients and normal participants were merged. Three clusters were presented in two cohorts, and survival analysis for a total of six groups and the presence or absence of chronic disease were confirmed. A confusion matrix-based statistical process was provided for survival and chronic disease for each of the three clusters based on DNA methylation.

**Methods**

**Data acquisition of TCGA**

The overall process of this study is presented in Fig. 1. To download the DNA methylation dataset analyzed in TCGA, the GDCquery function included in R's TCGAbiolinks package was used [18]. The R version used was 4.1.2. The parameters used when using the GDCquery function are as follows: The project used the project names for each of the 11 cancer types shown in Fig. 1 including TCGA-BRCA; data.category is "DNA Methylation"; data.type is "Methylation Beta Value"; workflow.type used "Liftover". Samples using Infinium HumanMethylation450 BeadChip (Infinium 450k chip) were used, and the number of male and female samples is shown in Fig. 1.
2.2 Data acquisition of KoGES

The epidemiologic and DNA methylation analysis dataset used in this study was obtained from the Ansung-Ansan cohort of KoGES surveyed between 2001 and 2002. Participants were between 40 and 69 years of age and resided in Ansung-si or Ansan-si, Gyeonggi-do, Republic of Korea. All participants provided written informed consent. For the DNA methylation dataset, a matrix for 446 people provided by KoGES was used, and the bioinformatic analysis process was carried out as described in previous studies [9, 10, 15]. This study was conducted in accordance with the Declaration of Helsinki with the approval of Korea University IRB (Institutional Review Board) (approval number: KUIRB-2020-0191-01).

Select dataset, integrate, and clustering

To correct for the batch effect on the DNA methylation dataset obtained from two different cohorts, the “ComBat” function included in the “sva” package of R was used [19]. To avoid bias due to differences in DNA methylation between sexes, the two sexes were analyzed separately [9]. The integration of the two datasets used the R default “merge” function, and the index was the probe ID used in the Infinium 450k chip. To annotate the gene symbol corresponding to each probe ID, the “getAnnotation” function of the “IlluminaHumanMethylation450kanno.ilmn12.hg19” package was used.

Each DNA methylation level for the sample was visualized and clustered using R's pheatmap package. To designate the \( k \) value in the \( k \)-means clustering algorithm, the value of “cutree_col” among the parameters of the “pheatmap” function was set to 3.

Visualization

For TCGA-derived carcinoma, follow-up date and mortality data were included, and a Kaplan-Meier (K-M) survival plot was drawn based on this data. Survival analysis was performed using the “survfit” function in the survival package of R, and a K-M plot was presented using the “ggsurvplot” function in the survminer package.

Results

Clustering DNA methylation levels

For a total of 11 carcinomas, a total of 19 clustering analyzes were performed by sex. For each result, DNA methylation level and cluster were provided simultaneously as a heatmap. Survival analysis in TCGA was presented, and a confusion matrix according to chronic disease was presented.

Among them, clustering results were presented for samples derived from 67 women from TCGA-PAAD and from 220 women from KoGES (Fig. 2). 26 DNA methylation analysis results satisfying standard deviation (SD) > 0.22 for each CpG site were presented as a heatmap and three clusters. Among the 287 samples, the number of samples corresponding to clusters 1, 2, and 3 was 93, 25, and 184 (Fig. 2A).
The results of clustering of samples from 141 females of TCGA-HNSC and 220 females of KoGES were presented (Fig. 3). By DNA methylation analysis, 19 CpG sites satisfying SD > 0.44 were presented as a heatmap and three clusters. Of the total 361 samples, 152, 117, and 92 were classified as clusters 1, 2, and 3.

The clustering of samples from 157 males of TCGA-COAD and 226 males of KoGES were presented (Fig. 4). The results satisfying SD > 0.25 for 10 CpG sites were presented as a heatmap and three clusters. Of the total 383 samples, 192, 129, and 62 were classified as clusters 1, 2, and 3.

The results of clustering of samples from 95 females of TCGA-LUSC and 220 females of KoGES were presented (Fig. 5). The results of 13 CpG sites by DNA methylation analysis satisfying SD > 0.25 were presented as a heatmap and three clusters. Of the total 315 samples, 112, 116, and 87 were classified as clusters 1, 2, and 3.

The results of clustering of samples from 270 male samples from TCGA-LUSC and 226 male samples from KoGES were presented (Fig. 6). Eight DNA methylation analysis results satisfying SD > 0.29 for each CpG site were presented as a heatmap and three clusters. 371, 44, and 81 of a total of 496 samples were classified as clusters 1, 2, and 3. In the case of 226 samples of KoGES, only cluster 1 was classified.

**Survival analysis**

Survival analysis was performed on clusters 1 and 3 in female samples of TCGA-PAAD, and 10 patients in cluster 1 had a higher 5-year follow-up survival rate than 54 patients in group 3 (Fig. 2B). Survival analysis was performed on clusters 1 and 3 in female samples of TCGA-HNSC, and 79 patients in cluster 1 had a higher 5-year follow-up survival rate than 184 patients in cluster 3 (Fig. 3B). Survival analysis was performed for clusters 1 and 3 in male samples of TCGA-COAD, and all 13 patients included in cluster 3 survived follow-up. On the other hand, about half of the 41 patients included in cluster 1 survived for about 2 years (Fig. 4B). Survival analysis was performed for the entire cluster in female samples of TCGA-LUSC. As a result, 7 samples belonging to cluster 1 showed the lowest survival rate (Fig. 5B). Survival analysis was performed on the whole cluster in male samples of TCGA-LUSC. As a result, a relatively low survival rate was confirmed in 137 samples belonging to cluster 1 (Fig. 6B).

**Confusion matrix analysis**

Chronic disease was checked for KoGES samples belonging to the same cluster as TCGA-PAAD. Among the 13 chronic diseases, the odds ratio was 2.18 between clusters 1 and 3 for factors related to diabetes (Fig. 2). This means that the prevalence of diabetes in the general population included in cluster 3 is higher. In the survival analysis of TCGA, the patients belonging to cluster 3 had a poor survival rate, so the pattern corresponding to cluster 3 can be considered a worse pattern than that of cluster 1. In female samples of KoGES, which belonged to the same cluster as TCGA-HNSC, the presence of chronic disease in clusters 1 and 3 was confirmed. The odds ratio for allergic disease was 1.23. In Cluster 3, the allergic disease rate was high, but there was no statistical significance (Fig. 3).

**Commonly related genes by venn diagram of five analysis**
The venn diagrams for the genes detected in the five analyzes are presented (Fig. 7). Commonly detected genes were listed as Table 1, and accession numbers were provided when the CpG sites were not located in the gene regions. A total of six CpG sites were commonly detected in the four analyzes, and all were analyzed in females. Nine CpG sites were detected in three assays, and all were analyzed in females, except for LUSC male assays. In female analysis of PAAD, a total of eight CpG sites were specifically identified.

<table>
<thead>
<tr>
<th>Names</th>
<th>No</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>COAD_F HNSC_F LUSC_F PAAD_F</td>
<td>3</td>
<td>AFAP1,NINJ2,HOOK2</td>
</tr>
<tr>
<td>HNSC_F LUSC_F LUSC_M PAAD_F</td>
<td>2</td>
<td>cg12036633,cg16814680</td>
</tr>
<tr>
<td>COAD_F LUSC_F LUSC_M PAAD_F</td>
<td>1</td>
<td>cg02159489</td>
</tr>
<tr>
<td>COAD_F HNSC_F PAAD_F</td>
<td>2</td>
<td>cg01156747,cg05528899</td>
</tr>
<tr>
<td>HNSC_F LUSC_F PAAD_F</td>
<td>6</td>
<td>MGMT,MRI1,UBAP2L,MYH15,cg27586797,cg21139150</td>
</tr>
<tr>
<td>HNSC_F LUSC_F LUSC_M</td>
<td>1</td>
<td>GNAS</td>
</tr>
<tr>
<td>HNSC_F PAAD_F</td>
<td>2</td>
<td>FBXL16,JRK</td>
</tr>
<tr>
<td>PAAD_F</td>
<td>8</td>
<td>RNF39,CCS,cg19214707,cg08103988,HLA-DRB1,HCG9,cg08779649,S100A14</td>
</tr>
<tr>
<td>HNSC_F</td>
<td>3</td>
<td>cg00456343,GDAP2,PMS2L11</td>
</tr>
<tr>
<td>COAD_F</td>
<td>2</td>
<td>NTNG1,PODN</td>
</tr>
<tr>
<td>LUSC_M</td>
<td>4</td>
<td>BCAT1,SOX2OT,KRT7,cg24000814</td>
</tr>
</tbody>
</table>

**Discussion**

Currently, cohort-based DNA methylation data have been accumulated from various samples, such as bloods of normal participants, bloods of cancer patients, and cancer tissues. For one individual, a comparative study between baseline and follow-up according to lifestyle was performed, and analysis methods were presented \[5, 15\]. In this study, DNA methylation data of TCGA cancer patients and healthy normal subjects obtained from KoGES were merged. I tried to find insights for prediction factors of cancer prognosis and chronic diseases from DNA methylation dataset, so I provided novel methods of DNA methylation data by merging and clustering.

After the multi-omics data of TCGA was published, modeling approaches have been conducted to classify normal and cancer, or each subtype of cancer \[20–22\]. Colorectal cancer susceptibility using 845
DNA methylation data GSE51032 of “the Italian arm of the European Prospective Investigation into Cancer and Nutrition (EPIC-Italy)”, a cohort study conducted in Italy, was presented. Comparative verification was performed in TCGA based on the results of GSE51032 [23]. Based on the results of TCGA-BRCA DNA methylation analysis, GO terms and pathways of 267 differentially methylated regions (DMRs) were discovered. For high- and low-risk classification, a model based on the DNA methylation level of QRFP, S100A16, TDRD1, and SMO genes was presented. Negative correlations for 10 genes were presented, and survival rates were analyzed for patients classified as high- and low-risk [24]. By these studies, DNA methylation patterns of cancer those are different from normal subjects were revealed.

The meta-analysis of cancer markers related to DNA methylation was performed. Koch et al. reported that 1,800 markers were presented in 14,743 research articles in 2018 [25], but only 14 of them were actually used commercially. Each gene was characterized as a biomarker based on DNA methylation for GSTP1, APC, RASSF1, NDRG2, BMP3, SEPT9, SHOX2, TWIST1, OTX1, ONECUT2, MGMT, BCAT1, IKZF1 genes. In this study, a Venn diagram was presented for genes detected at least once in a total of five analyzed carcinomas (Fig. 7), and MGMT and BCAT1 genes were identified in common. MGMT gene was detected in female HNSC, LUSC, and PAAD samples to classify prognosis as well as chronic diseases.

In four female carcinoma samples (COAD, HNSC, LUSC, and PAAD), three genes (AFAP1, NINJ2, and HOOK2) were commonly identified as classifier, and overexpression of the three genes was related to oncogenesis. AFAP1 gene were hypermethylated in poor cluster than good cluster. In the antisense of AFAP1 gene, AFAP1-AS1 gene has been known as an oncogenic long non-coding RNA in human carcinoma. AFAP1-AS1 was upregulated in lung cancer, and the function was revealed to promote invasion and metastasis [26]. High expression of IncRNA AFAP1-AS1 promotes the progression of colon cancer and predicts poor prognosis [27].

NINJ2 gene codes Ninjurin 2 protein, and its overexpression promotes human colorectal cancer cell growth in vitro and in vivo [28]. In NINJ2 gene, rs118050317 polymorphism was related to endometrial cancer risk [29]. NINJ2 overexpression promotes glioma cell growth [30].

The limitations of this study are as follows. First, TCGA and KoGES samples were analyzed separately according to sex, but age was not separately analyzed. Because carcinoma genesis is detected after the age of 50, the age ranges (40–60 years old) are similarly matched. Second, the results were not confirmed by the experiment. Third, TCGA and KoGES samples were derived from different sources except for TCGA-LAML. In the further studies, cross-validation experiments is needed in normal-cancer matched samples.

In each country, the general public omics datasets will be accumulated and continue to be released including KoGES. By integrating omics datasets in an appropriate way, the value of omics datasets can be maximized to discover disease-related factors.

Conclusions
In this study, a method for integrated analysis of public DNA methylation data was presented. Using the methodology of this study, CpG sites for cancer research can be extracted from TCGA, and chronic disease-related factors can be extracted from cohorts for normal subjects such as KoGES. The results of clustering in this study will be very helpful in predicting the patient's prognosis and suggesting health management directions.

**Declarations**

**Author contributions**

J-AG contributed to conception, methodology, software, design, data analysis, investigation, data curation, funding acquisition, revising and writing the manuscript and finalize it.

**Funding Sources**

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI21C0012) and the National Research Foundation (NRF) funded by the Ministry of Education (grant number: NRF-2020R1I1A1A01052701).

**Competing of Interest Statement**

The author has no conflicts of interest regarding this study.

**Data availability Statement**

The source code used in this manuscript will be provided upon request.

**Acknowledgements**

None.

**Consent for the Publication**

Not applicable.

**Statement of Ethics**

This study was conducted in accordance with the Declaration of Helsinki with the approval of Korea University IRB (Institutional Review Board) (approval number: KUIRB-2020-0191-01).

**References**


15. Gim J-A: **Integrative approaches of DNA methylation patterns according to age, sex, and longitudinal changes.** 2022.


Figures
Processes of this study. The number of samples analyzed by 450k DNA methylation was presented separately for KoGES normal participants and two sexes of each cancer type. DNA methylation information between normal participants and cancer patients was merged, and after clustering, genes with large standard deviations for each gene were selected.

Figure 2

Clustering, survival analysis and confusion matrix of female DNA methylation. (A) The clustering results of samples from 67 females of TCGA-PAAD and 220 females of KoGES were presented as heatmaps. (B) Kaplan-Meier plots for TCGA-PAAD patients in Clusters 1 and 3. (C) Confusion matrix according to diabetes status for 220 women in KoGES.
Figure 3

Clustering, survival analysis and confusion matrix of female DNA methylation. (A) The clustering results of 141 female samples from TCGA-HNSC and 220 female samples from KoGES are presented as heatmaps. (B) Kaplan-Meier plots for TCGA-HNSC patients in Clusters 1 and 3. (C) Confusion matrix according to allergic diseases in 220 women of KoGES
Clustering, survival analysis and confusion matrix of male DNA methylation. (A) The clustering results of 157 male samples from TCGA-COAD and 226 male samples from KoGES are presented as heatmaps. (B) Kaplan-Meier plots for TCGA-COAD patients in Clusters 1 and 3. (C) Confusion matrix according to the presence of gastritis or stomach ulcer in 226 men of KoGES.
Figure 5

Clustering, survival analysis and confusion matrix of female DNA methylation. (A) The clustering results of samples from 95 females of TCGA-LUSC and 220 females of KoGES are presented as heatmaps. (B) Kaplan-Meier plots for TCGA-LUSC patients in Clusters 1, 2 and 3. (C) Confusion matrix according to the presence or absence of asthma in 220 women of KoGES

![Figure 5](image)

Figure 6

Clustering, survival analysis and confusion matrix of male DNA methylation. (A) The clustering results of 270 male samples from TCGA-LUSC and 226 male samples from KoGES are presented as heatmaps. (B) Kaplan-Meier plots for TCGA-LUSC patients in Clusters 1, 2 and 3. The confusion matrix for 220 men in KoGES was classified by chronic disease, but clustering was not provided

![Figure 6](image)
Figure 7

Venn diagram of five analyses