scientific reports

Check for updates

OPEN

# Machine learning models for pancreatic cancer diagnosis based on microbiome markers from serum extracellular vesicles

Doeun Lee[1,5], Chanhee Lee[1,5], Kyulhee Han[1], Taewan Goo[1], Boram Kim[1], Youngmin Han[2], Wooil Kwon[2], Seungyeoun Lee[3], Jin-Young Jang[2]✉ & Taesung Park[1,4]✉

**Pancreatic cancer (PC) is a fatal disease with an extremely low 5-year survival rate, mainly because of its poor detection rate in early stages. Given emerging evidence of the relationship between microbiota composition and diseases, this study aims to identify microbiome markers linked to the diagnosis of pancreatic cancer. We utilized extracellular vesicles (EVs) data obtained from blood samples of 38 pancreatic cancer patients and 51 health controls. Least absolute shrinkage and selection operator (LASSO) and stepwise method were used to obtain some candidate markers in genus and phylum levels. These markers were used to develop various machine learning models including logistic regression (LR), random forest (RF), support vector machine (SVM), and Deep Neural Network (DNN) methods. In phylum level, DNN performed best with three markers (*Verrucomicrobia*, *Actinobacteria* and *Proteobacteria*) selected by stepwise method with the test AUC 0.959. In genus level, DNN using 11 markers selected by LASSO (*Ruminococcaceae UCG-013*, *Ruminiclostridium*, *Propionibacterium*, *Lachnospiraceae NK4A136 group*, *Corynebacterium.1*, *Akkermansia*, *Mucispirillum*, *Pseudomonas*, *Diaphorobacter*, *Clostridium sensu stricto 1* and *Turicibacter)* outperformed others with 0.961 test AUCs. These results highlight the potential of microbiome markers and prediction models in clinical studies of PC diagnosis.**

**Keywords** Pancreatic cancer, Microbial extracellular vesicles, Microbiome markers, Early diagnosis, Machine learning

Pancreatic cancer (PC) is a highly aggressive malignancy with a dismal prognosis, primarily due to the lack of effective screening methods and the asymptomatic nature of early-stage disease, which often results in late-stage diagnosis[1]. The five-year survival rate remains below 8%, highlighting the urgent need for improved early detection strategies[2]. The poor treatment response is largely attributed to the tumor microenvironment, which fosters an immunosuppressive landscape and contributes to resistance against conventional chemotherapy[3]. Additionally, the dense stromal composition of pancreatic tumors creates a physical barrier that limits drug penetration, further reducing therapeutic efficacy[4]. Early diagnosis plays a pivotal role in expanding treatment options, particularly by increasing the likelihood of surgical resection, which remains the only curative approach[5]. Even with advancements in targeted therapies and immunotherapy, early detection remains crucial for improving patient outcomes by enabling timely intervention before disease progression to advanced, unresectable stages[6].

Studies have clearly shown that the human microbiota plays a key role in human health and disease, substantiating its potential use as a predictor for various diseases including cancers[7–9]. Microbiome profiles identified in fecal, urine, or serum samples of cancer patients have been utilized not only to help determine the course of cancer progression but also to suggest risk prediction models. With advances in sophisticated analytical methods using machine learning algorithms, it has become feasible to decipher the functional relationships between a specific microbiome and tumor progression[10]. Several studies have implicated candidates for early diagnostic biomarkers related to high risk of cancers including colon, hepatocellular carcinoma and breast

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea. [2]Department of Surgery and Cancer Research Institute, Seoul National University College of Medicine, Seoul 03080, Korea. [3]Department of Applied Mathematics, Sejong University, Seoul 03080, Korea. [4]Department of Statistics, Seoul National University, Seoul 08826, Korea. [5]Doeun Lee and Chanhee Lee contributed equally to this work. ✉email: jangjy4@snu.ac.kr; tspark@stats.snu.ac.kr

cancer[11]. Recently, the fecal microbiome signature of pancreatic adenocarcinoma was obtained by analyzing a cohort composed of 30 patients with PC and 29 controls of healthy subjects and patients with non-alcoholic fatty liver disease[12]. Fourteen bacterial features discriminated between PC and controls via a random forest (RF) model with an AUC of 0.825. The gut microbial profiles of PC in China have been reported in a cohort of 85 PC and 57 matched healthy controls by MiSeq sequencing Gut microbial profile analysis by MiSeq sequencing of pancreatic carcinoma patients in China[13]. The results showed that microbial markers based on 40 genera associated with PC achieved a high classification power, with AUC of 0.842.

Microbial information obtained from extracellular vesicles (EVs) excreted from bacteria can serve as a stable and enriched source of microbial biomarkers, enabling the assessment of microbiota composition and its role in disease. EVs circulate in the bloodstream, providing a non-invasive means to capture microbial signatures from distant organs, including the tumor microenvironment[14]. Their stability in biological fluids and ability to cross physiological barriers make them particularly valuable for their potential applications in early cancer detection[15,16]. In the context of pancreatic cancer diagnosis, EV-derived microbial markers present a promising non-invasive alternative to conventional sampling methods, providing a more comprehensive and systemic representation of disease-associated microbiome alterations. Recent studies have analyzed diagnostic models of various cancer types such as glioblastoma[17], colorectal cancer[18], and ovarian cancer[19] using microbial data from EVs. However, PC has not been widely studied in this regard.

Our recent microbiome profiling study using serum EVs identified biomarkers indicating PC and allowed us to propose a diagnosis model classified in terms of operational taxonomic units (OTUs)[20]. Candidate markers were first selected using nine single marker selection methods: microbiome specific methods such as metagenome-seq (gaussian and log-normal)[21], ZIBSeq[22], ANCOM[23] and CLR permutation[24], simple statistical analysis such as Wilcoxon rank sum test[25], and RNA-seq data specific methods such as DESeq2 (LRT, Wald)[26] and edgeR[23,27]. For those markers commonly identified by these methods, the best marker combination was selected by exhaustive search. Logistic regression (LR) models using these markers yielded high area under the curve (AUC) of receiver operating characteristics (ROC) of 0.966 and 0.913 in phylum and genus level in the test set, respectively.

While this model was shown to provide a reasonably well-predicted performance, there remains considerable room for improvement. For example, single marker selection does not consider the well-known, strong correlations among microbiome markers. In building prediction models, it is more efficient to choose, as much as possible, those markers less correlated and more complementary. Furthermore, many available machine learning models may provide better performance than LR.

In this study, we adopt multiple marker selection methods to account for the multi-collinearity among the candidate markers. Moreover, we apply machine learning models such as random forest (RF), support vector machine (SVM), and deep neural network (DNN) to achieve better prediction performance. RF was selected for its robustness in handling high-dimensional data and mitigating overfitting through ensemble learning. SVM was chosen for its ability to classify non-linearly separable data using various kernel functions. DNN was employed to capture complex hierarchical patterns in microbiome data, potentially enhancing prediction accuracy. These models were selected to comprehensively evaluate different classification approaches and identify the most effective method for pancreatic cancer diagnosis. The performance of each combination of marker selection and machine learning model is evaluated to choose the optimal combination. At phylum level, a DNN model using 3 markers (*Verrucomicrobia*, *Actinobacteria* and *Proteobacteria*) was selected with test AUC of 0.959. In genus level, a DNN model using 11 markers (*Ruminococcaceae UCG-013*, *Ruminiclostridium*, *Propionibacterium*, *Lachnospiraceae NK4A136 group*, *Corynebacterium.1*, *Akkermansia*, *Mucispirillum*, *Pseudomonas*, *Diaphorobacter*, *Clostridium sensu stricto 1* and *Turicibacter)* yielded the highest test AUC with 0.961.

This paper is organized as follows: the materials and methods section introduce the marker selection methods and prediction models based on machine learning methods used for PC prediction. The marker sets selected via each method and the performance of the models are listed in the results section.

## Materials and methods

### Study population

The information on PC patients was obtained from Seoul National University Hospital, and data of healthy controls were received from Boramae Medical Center of Seoul National University Hospital and Inje University Haeundae Paik Hospital. The patients included in this analysis were diagnosed and underwent surgical resection between 2009 and 2015. For healthy controls, people with no history of any cancers, including PC, were considered. Patients of all cohorts who had clinical characteristics (age and sex) were included in our study, with their data anonymized. To mitigate potential confounding effects arising from differences in clinical characteristics, we applied propensity score matching (PSM), specifically controlling for age and gender. This approach ensured that the case and control groups were balanced with respect to these covariates, thereby reducing biases and improving the validity of our findings. Our study was approved by the Institutional Review Board, and all methods were carried out in accordance with relevant guidelines and regulations.

The 16s rRNA data was obtained from blood samples. The gg_13_5_99 GREENGENES reference was used in taxonomic assignment, and UCLUST and QIIME were performed[28,29]. The process of data preparation and microbiome sequencing is detailed in our previous research[20]. From the resulting OTU table, samples with low sequence count (< 2500) and OTUs constituting less than 0.005% of the total count were filtered out. The data were then subjected to centered log ratio (CLR) transformation to compare the compositional differences of PC patients and healthy controls.

To minimize the bias of the covariate effects in the case and control groups, we adopted the PSM procedure[30], a statistical matching method which minimizes the bias induced by the confounding variable, thus reducing the

confounding effect of the data in observational studies[31]. The application of PSM to microbiome data aims to remove the effect of specific characteristics and obtain the pure effect of the target microbiome. As our data were highly heterogeneous in age and sex distribution between PC and healthy groups, PSM was applied to control the confounding effects caused by these covariates.

For model building, we divided the dataset into model development (MD) and test sets in a 1:1 ratio. This equal partitioning was maintained for consistency with our previous study, which employed a similar analysis pipeline using logistic regression (LR)[20]. Since this study extends the prior work by incorporating machine learning (ML) models, we retained the same partitioning strategy to ensure methodological consistency and comparability. Additionally, given the relatively small sample size after propensity score matching ($n = 90$; PC patients = 38, controls = 52), this approach ensured a sufficiently large independent test set for robust model evaluation. The MD set was further divided into a training set and validation set at the same proportion, stratified by the healthy controls and PC patients for twofold cross-validation. The AUC value of the validation set was used as a criterion for selecting the model. We selected candidate markers using a multiple marker selection method based on LASSO analysis and stepwise regression. These markers were then used to develop prediction models.

## Multiple marker selection methods

As mentioned, LASSO regression and stepwise method were used to select candidate markers. All the selection processes were performed using the MD set with twofold cross-validation, the marker set with the biggest average validation AUC being selected.

In the LASSO regression, 100 $\lambda$ values were obtained, and the searching range of optimal lambda was set between $\lambda_{min}$ and $\lambda_{1se}$ to avoid overfitting[32]. We compared the validation AUC value of the fitted model in this predetermined range and selected the $\lambda$ value yielding the highest validation AUC, the corresponding markers then being utilized.

For the stepwise method, two approaches were applied. In the first approach (stepwise1), the model that yielded the least AIC was selected, and the marker combination used for the model was then selected in a forward-backward manner. The performance was averaged and the best combination (yielding the maximum validation performance) became the candidate combination to be tested further.

The second approach (stepwise2) first determined the number of markers. After using the stepwise method to fix the optimal number of markers, we used an exhaustive search to select those variables yielding the best validation AUC. The variables were added one by one until the number of markers reached the number that had been set.

In addition to those markers found by the selection methods explained above, we also considered three more marker sets for comparison: (1) whole markers, (2) only clinical information (age and sex) without markers, and (3) markers selected through single marker selection. The candidate markers were searched in both phylum and genus levels and then utilized to develop prediction models using LR, RF, SVM and DNN. The detailed model building procedure is given below.

## Prediction model

We constructed predictive models identifying PC patients from the healthy control group, with markers selected via the methods illustrated above and clinical information including age and sex. The models we considered were LR, RF, SVM and DNN. All models except LR require optimized hyper-parameters to get the best performance (Supplementary Table S1). To determine the optimal hyperparameters, we employed a grid search approach, systematically evaluating combinations of predefined parameter values to identify the configuration that maximized model performance. By using twofold cross-validation, an optimal model yielding the highest average value of validation AUC was chosen.

RF utilizes many classifiers to learn. We set the number of the subtrees with values (4, 8, 16, 32, 64), and two classification criteria, gini and entropy. To prevent overfitting, the number of subtrees was limited to 100 considering the sample size.

For the SVM model, four different kernels (linear, polynomial, RBF and sigmoid) were used with each candidate hyper-parameter value. The cost, degree, and gamma value needed to be optimized, and the range of the hyper-parameters were set with values (0.5, 0.8, 1, 2, 4, 8, 16, 20, 25, 30), (2, 3, 4) and (0.1, 5, 10), respectively.

A DNN model was also implemented for prediction. The number of hidden units in each layer was determined considering the size of the input feature of the previous layer. We calculated the hidden units of the current layer as (0.2, 0.4, 0.6, 0.8), the proportion of number of hidden units in the previous layer, combined with the set of layer numbers (3, 4) and the batch size (6, 8, 10). Exceptionally, a (0.5, 1) hidden unit proportion was used for the null model since it has only two features. The models were trained with Adam optimizer[33] and binary cross-entropy loss function, and all layers except the output layer used a rectified linear unit (RELU) activation function at learning rate 0.0001. The sigmoid activation function was used for the output layer. Also, the early stopping method[34] was introduced with 100 patience. Since the sample size for the model was much smaller than the feature number, the data were standardized and cross-validated to avoid over-fitting. Some models with a small feature size skipped some hyper-parameter combinations like hidden unit proportion 0.2 to avoid a number of hidden units smaller than 1. Using average validation AUC value as a criterion, the best combination was selected for each marker set. The optimized model was fitted with the whole MD set, and the final performance then calculated with the test set.

The RF and DNN models were performed in Python using Keras (Version 2.4.3, https://github.com/keras-team/keras) and the Tensorflow (Version 2.3.0, https://github.com/tensorflow/tensorflow) library. SVM analysis was also performed using R version 1.2.5 on Window 10 (Version 1.2.5, http://www.R-project.org).

## Results

### Sample selection and bias mitigation using propensity score matching

Among 234 samples with clinical data, 198 were selected by excluding samples with low sequence count (< 2500). To reduce the bias induced by clinical information, the remaining data were filtered with propensity score matching (PSM) to yield 90 samples of 38 PC patients (21 females, 17 males) and 52 healthy controls (35 females, 17 males). From 1134 OTUs in the samples, 20 phyla and 296 genera were assigned. The distributions of age and sex did not show any significant differences between PC and health groups: $p = 0.35$ by the chi-square test for sex, and $p = 0.51$ by Wilcoxon rank-sum test for age (Fig. 1).

### Selecting candidate biomarkers using multiple marker selection methods

The markers related to PC were selected using least absolute shrinkage and selection operator (LASSO)[35] and two stepwise[36] methods. The markers were selected at both the phylum and genus levels to capture microbial community characteristics at different taxonomic resolutions. Phylum-level analysis provides a broad overview of microbial shifts, identifying high-level taxonomic changes that may indicate dysbiosis, while genus-level markers offer more specific insights into particular microbial taxa associated with pancreatic cancer. This combined approach allows for a more comprehensive characterization of the microbiome's role in disease progression and enhances the robustness of biomarker selection.

The result of LASSO regression in phylum level is tabulated in Table 1A. Instead of presenting one optimal value of $\lambda$, we listed the markers according to various $\lambda$ values within the range of $\lambda_{1se}$ and $\lambda_{min}$ to avoid overfitting (shaded in yellow). We calculated the average AUCs of the training and validation sets using twofold cross-validation for each $\lambda$ value. The best fitted model was obtained when using $\lambda_{min}$ with AUC 0.855. Accordingly, ten markers, *Verrucomicrobia*, *Deferribacteres*, *Actinobacteria*, *Saccharibacteria*, *Armatimonadetes*, *Absconditabacteria (SR1)*, *Acidobacteria*, *Cyanobacteria*, *Euryarchaeota*, and *Deinococcus-Thermus* were selected and used for prediction model building. The average training and validation AUCs of the model using 100 $\lambda$ values and binomial deviance of LASSO analysis were also obtained (Supplementary Figure S1). In LASSO analysis in genus level shown in Table 1B, eleven markers were selected with AUC 0.936, the markers being: *Ruminococcaceae UCG-013*, *Ruminiclostridium*, *Propionibacterium*, *Lachnospiraceae NK4A136 group*, *Corynebacterium 1*, *Akkermansia*, *Mucispirillum*, *Pseudomonas*, *Diaphorobacter*, *Clostridium sensu stricto 1*, and *Turicibacter*. The model performance using 100 $\lambda$ values and the binomial deviance of the model were also calculated (Supplementary Figure S2).

In the stepwise1 method for phylum level marker selection, markers were not removed in backward elimination. In this process, model obtained the highest validation AUC of 0.930 with *Verrucomicrobia*, *Actinobacteria*, and *Proteobacteria*. The performance of the model by each step is shown in Table 2A, the optimal marker set being shaded in yellow. In stepwise1 method for genus level marker selection, *Akkermansia*, *Propionibacterium*, and *Burkholderia-Paraburkholderia* were selected with AUC 0.836. The results when each marker was added one by one are summarized in Table 2B.

In the stepwise2 method for phylum level marker selection, *Proteobacteria* was replaced with *Deferribacteres*, resulting in selecting *Verrucomicrobia*, *Actinobacteria* and *Deferribacteres* as the final marker set, achieving 0.964 AUC. The performance of each step is listed in Table 3A. However, different markers were selected when applying the stepwise2 method. The *Lachnospiraceae NK4A136* group, *Ruminococcaceae UCG-013*, and *Bacteroides* were selected with AUC of 0.915, the performance of each marker being listed in Table 3B. No markers were removed during the backward stepwise process in either method.

Compared with the single marker selection results[13], some overlapping markers were selected in both methods, showing consistency. Additionally, our current study provided some new candidate biomarkers for developing the performance of PC prediction models (Supplementary Table S6). The markers in the "Single" column in Supplementary Table S6 were referenced in our previous analysis[13]. Two phylum-level markers, *Actinobacteria* and *Verrucomicrobia*, were selected in all methods. In genus level, there were no commonly chosen biomarkers in all methods, but *Propionibacterium*, *Akkermansia*, *Ruminococcaceae UCG-013* and *Lachnospiraceae NK4A136*

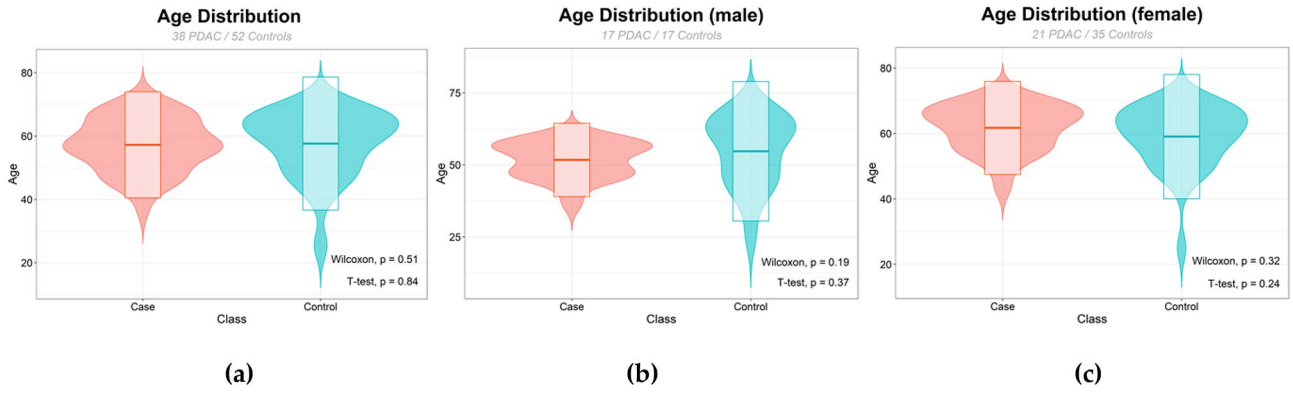


**Fig. 1**. Age distribution of the samples and p-value of Wilcoxon and t-test: (**a**) Age distribution of all samples; (**b**) Age distribution of males; (**c**) Age distribution of females.

| (A) Phylum | | | | |
|---|---|---|---|---|
| **# markers** | **Lambda** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | 0.256802 | sex + age | 0.688 | 0.663 |
| 1 | 0.233989 | + *Verrucomicrobia* | 0.772 | 0.771 |
| 2 | 0.213202 | + *Deferribacteres* | 0.826 | 0.807 |
| 3 | 0.194261 | + *Actinobacteria* | 0.865 | 0.841 |
| 5 | 0.063612 | + *Saccharibacteria*<br>+ *Armatimonadetes* | 1.000 | 0.840 |
| 6 | 0.057961 | + *Absconditabacteria (SR1)* | 1.000 | 0.836 |
| 7 | 0.043845 | + *Acidobacteria* | 1.000 | 0.836 |
| 8 | 0.03995 | + *Cyanobacteria* | 1.000 | 0.836 |
| 9 | 0.036401 | + *Tenericutes* | 1.000 | 0.840 |
| 10 | 0.022861 | + *Euryarchaeota*<br>+ *Deinococcus-Thermus*<br>– *Tenericutes* | 1.000 | 0.855 |
| 9 | 0.015757 | - *Deferribacteres* | 1.000 | 0.863 |
| 10 | 0.013082 | + *Planctomycetes* | 1.000 | 0.871 |
| 11 | 0.01192 | + *Chloroflexi* | 1.000 | 0.878 |
| 10 | 0.002953 | - *Acidobacteria* | 1.000 | 0.894 |
| **(B) Genus** | | | | |
| **# markers** | **Lambda** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | 0.300095 | sex + age | 0.758 | 0.684 |
| 3 | 0.286455 | + *Ruminococcaceae UCG-013* + *Ruminiclostridium*<br>+ *Propionibacterium* | 0.783 | 0.699 |
| 4 | 0.227011 | + *Lachnospiraceae NK4A136 group* | 0.942 | 0.831 |
| 6 | 0.206844 | + *Corynebacterium 1*<br>+ *Akkermansia* | 0.975 | 0.871 |
| 7 | 0.179902 | + *Mucispirillum* | 0.991 | 0.900 |
| 9 | 0.16392 | + *Pseudomonas* + *Diaphorobacter* | 0.996 | 0.903 |
| 10 | 0.149358 | + *Clostridium sensu stricto 1* | 0.996 | 0.916 |
| 11 | 0.085468 | + *Turicibacter* | 1.000 | 0.936 |
| 12 | 0.074336 | + *Actinomyces* | 1.000 | 0.941 |
| 13 | 0.070957 | + *Campylobacter* | 1.000 | 0.941 |
| 14 | 0.067732 | + *Megamonas* | 1.000 | 0.944 |
| 16 | 0.05891 | + *Bacillus* + *Rubellimicrobium* | 1.000 | 0.937 |
| 17 | 0.056232 | + *Alistipes* | 1.000 | 0.929 |
| 18 | 0.053676 | + *Sutterella* | 1.000 | 0.932 |
| 17 | 0.035316 | + *Chromohalobacter* -*Turicibacter* - *Megamonas* | 1.000 | 0.937 |
| 18 | 0.027987 | + *Brevibacterium* | 1.000 | 0.932 |
| 17 | 0.021171 | - *Propionibacterium* | 1.000 | 0.929 |
| 18 | 0.016015 | + *Turicibacter* | 1.000 | 0.933 |
| 19 | 0.010537 | + *Cytophaga* | 1.000 | 0.925 |
| 20 | 0.004354 | + *Mycobacterium* | 1.000 | 0.921 |
| 21 | 0.00345 | + *Roseburia* | 1.000 | 0.921 |
| 22 | 0.003144 | + *Ruminococcaceae UCG-014* | 1.000 | 0.921 |

**Table 1**. Selected markers via the LASSO analysis in phylum and genus. The range between $\lambda_{1se}$ and $\lambda_{min}$ is shaded in yellow.

*group* were selected by two of them. Some species within *Actinobacteria*, such as certain *Bifidobacterium* strains, contribute to butyrate metabolism, although the primary butyrate producers belong to the *Firmicutes* phylum[37]. Butyrate is a short-chain fatty acid that plays a crucial role in modulating the immune response. It acts as an anti-inflammatory agent by inhibiting nuclear factor-kappa B (NF-κB) activation in colon epithelial cells[38], thereby promoting gut homeostasis and reducing chronic inflammation. Given that chronic inflammation is a key factor in tumor development, the observed decrease in *Actinobacteria* abundance in PC patients suggests a potential loss of protective anti-inflammatory effects, which may contribute to a tumor-promoting environment[39].

In contrast, *Verrucomicrobia*, particularly *Akkermansia muciniphila*, was found to be more abundant in PC patients than in healthy controls. *Akkermansia* is known for its role in maintaining glucose homeostasis and strengthening gut barrier integrity[40]; however, its relationship with pancreatic cancer remains insufficiently studied. Emerging evidence suggests that altered glucose metabolism is a hallmark of cancer, and the increased abundance of *Akkermansia* may reflect meta bolic shifts rather than a direct causal role in tumorigenesis[41].

| (A) Phylum | | | |
|---|---|---|---|
| **# markers** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | class ~ sex + age | 0.606 | 0.581 |
| 1 | + *Verrucomicrobia* | 0.983 | 0.907 |
| 2 | + *Actinobacteria* | 1.000 | 0.874 |
| 3 | + *Proteobacteria* | 1.000 | 0.930 |
| **(B) Genus** | | | |
| **# markers** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | class ~ sex + age | 0.606 | 0.581 |
| 1 | + *Akkermansia* | 0.985 | 0.825 |
| 2 | + *Propionibacterium* | 1.000 | 0.831 |
| 3 | + *Burkholderia-Paraburkholderia* | 1.000 | 0.836 |

**Table 2**. Selected markers using stepwise1 procedure in phylum and genus level. The mean train and validation AUCs are recorded.

| (A) Phylum | | | |
|---|---|---|---|
| **# markers** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | class ~ sex + age | 0.599 | 0.545 |
| 1 | + *Verrucomicrobia* | 0.960 | 0.931 |
| 2 | + *Actinobacteria* | 0.970 | 0.964 |
| 3 | + *Deferribacteres* | 0.972 | 0.964 |
| **(B) Genus** | | | |
| **# markers** | **marker** | **$AUC_{tr}$** | **$AUC_{val}$** |
| 0 | class ~ sex + age | 0.599 | 0.545 |
| 1 | + *Lachnospiraceae NK4A136 group* | 0.907 | 0.852 |
| 2 | + *Ruminococcaceae UCG-013* | 0.947 | 0.858 |
| 3 | + *Bacteroides* | 0.955 | 0.915 |

**Table 3**. Selected markers using stepwise2 procedure in phylum and genus level. The mean train and validation AUCs are recorded.

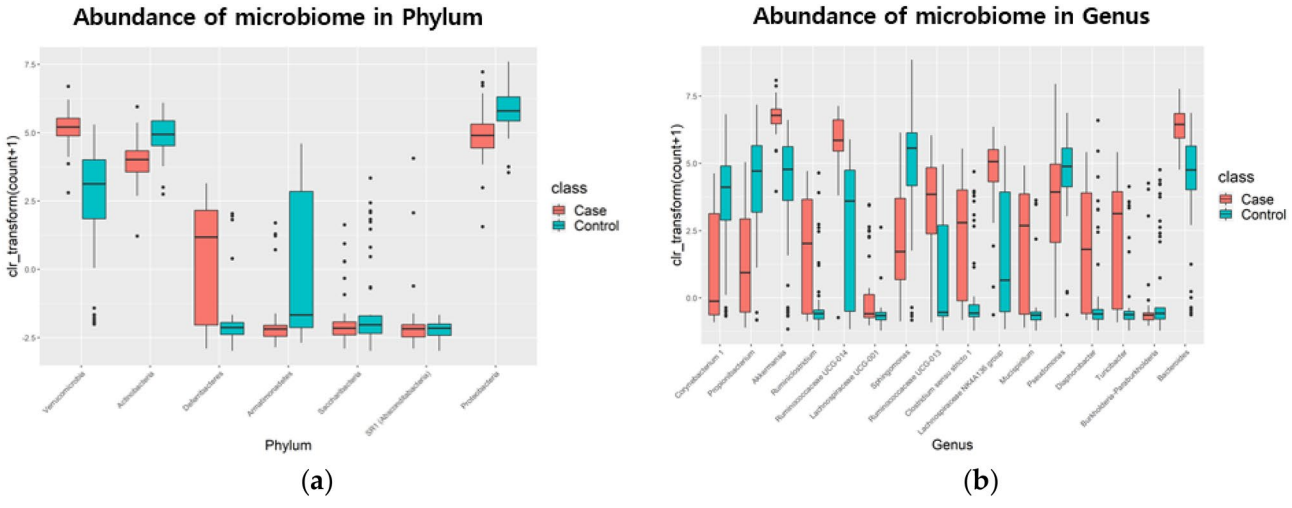


**Fig. 2**. Overall abundance of selected markers in phylum and genus level.

Further studies are required to elucidate whether these microbial shifts are drivers or consequences of pancreatic cancer progression. The overall abundance of selected phylum level markers in PC patients and the healthy group is shown in Fig. 2A and genus level in Fig. 2B.

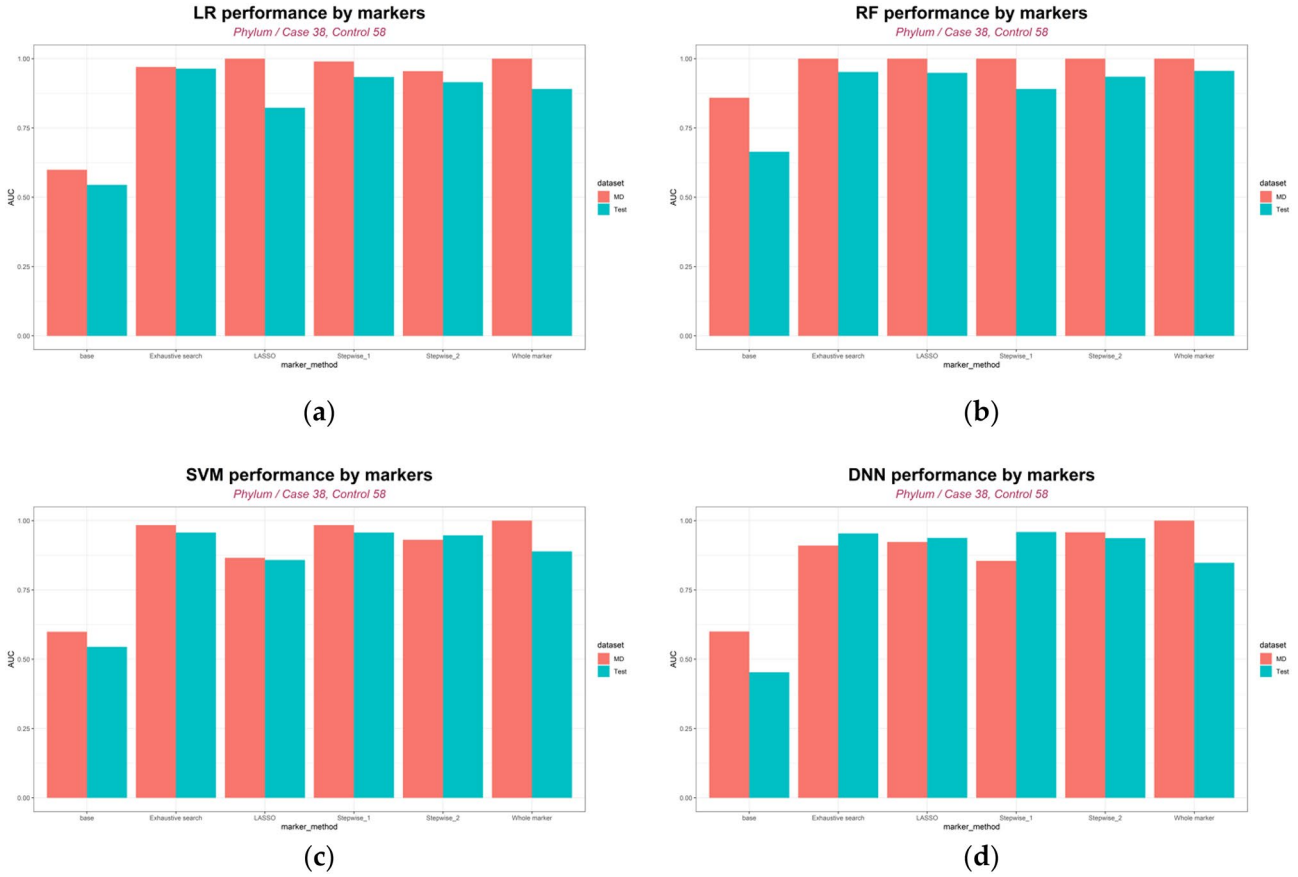


**Fig. 3**. Best model performance by marker set in phylum level for test data. (**a**) Performance of LR; (**b**) Performance of RF; (**c**) Performance of SVM; (**d**) Performance of DNN.

## Prediction model for early diagnosis of pancreatic cancer

In order to develop the prediction model for pancreatic cancer, we used several statistical models and machine learning techniques. The models we considered were LR, RF, SVM, and DNN. After each model was tuned with optimal hyper-parameters to obtain the highest value of average validation AUC, the test AUC and average model development (MD) AUC were calculated.

The performance of LR with various marker sets was calculated, as shown in Figs. 3A and 4A (Supplementary Table S2). In phylum level, the model with stepwise1 markers performed best, with 0.934 of AUC, followed by stepwise2 and LASSO with AUCs of 0.915 and 0.818, much higher than the case using whole markers (0.891 of AUC). In genus level, the model built with markers selected by stepwise1 yielded 0.955 of AUC. Models with stepwise2 and LASSO performed 0.924 and 0.896, respectively. In both cases, the models outperformed the case of using whole markers (full model) with AUC of 0.891 and 0.378, respectively.

For RF models selecting multiple markers, the hyper-parameter combination for best model was chosen. For all markers, (entropy, 64) was selected for (number of subtrees, classification criteria). The AUC was calculated with the optimal hyper-parameter (Supplementary Table S3), and the performance is shown in Figs. 3B and 4B. The best performance in phylum level was obtained from the model developed with markers selected via stepwise2, achieving test AUC of 0.935. The score was slightly higher than the following criteria, which is 0.922 of AUC with the LASSO method and 0.891 of AUC with the stepwise1 method. In genus level, markers selected with stepwise1 achieved the best performance, with 0.929 in AUC. This was followed by stepwise2, with test AUC 0.921 and LASSO with test AUC 0.877. In both phylum and genus level, the full model performed best, with test AUC of 0.956 and 0.957, respectively.

In the SVM model, various hyper-parameter sets were chosen but models using stepwise1 performed best in both levels. With hyper-parameter combination (kernel, cost, degree, gamma), the best model had an AUC of 0.957 with (linear, 0.5, default, 1) and 0.943 with (polynomial, 4, 2, default) in phylum and genus level, respectively. Compared with the full model (AUC of 0.899 and 0.980 in each level), the selected model outperformed in phylum level, but the full model performed better in genus level. The performance is shown in Figs. 3C and 4C.

The test AUC of the best performing DNN model was 0.959 and 0.961 in phylum and genus level, respectively (Supplementary Table S5). In phylum level, the model with stepwise1 markers, with (6, [0.6, 0.2, 0.6]) hyper-parameter combination for (batch size, hidden unit proportion) was selected, and a model using LASSO markers with (10, [0.8,0.8,0.8]) hyper-parameter combination was selected for the genus level. Here, the stepwise2 and

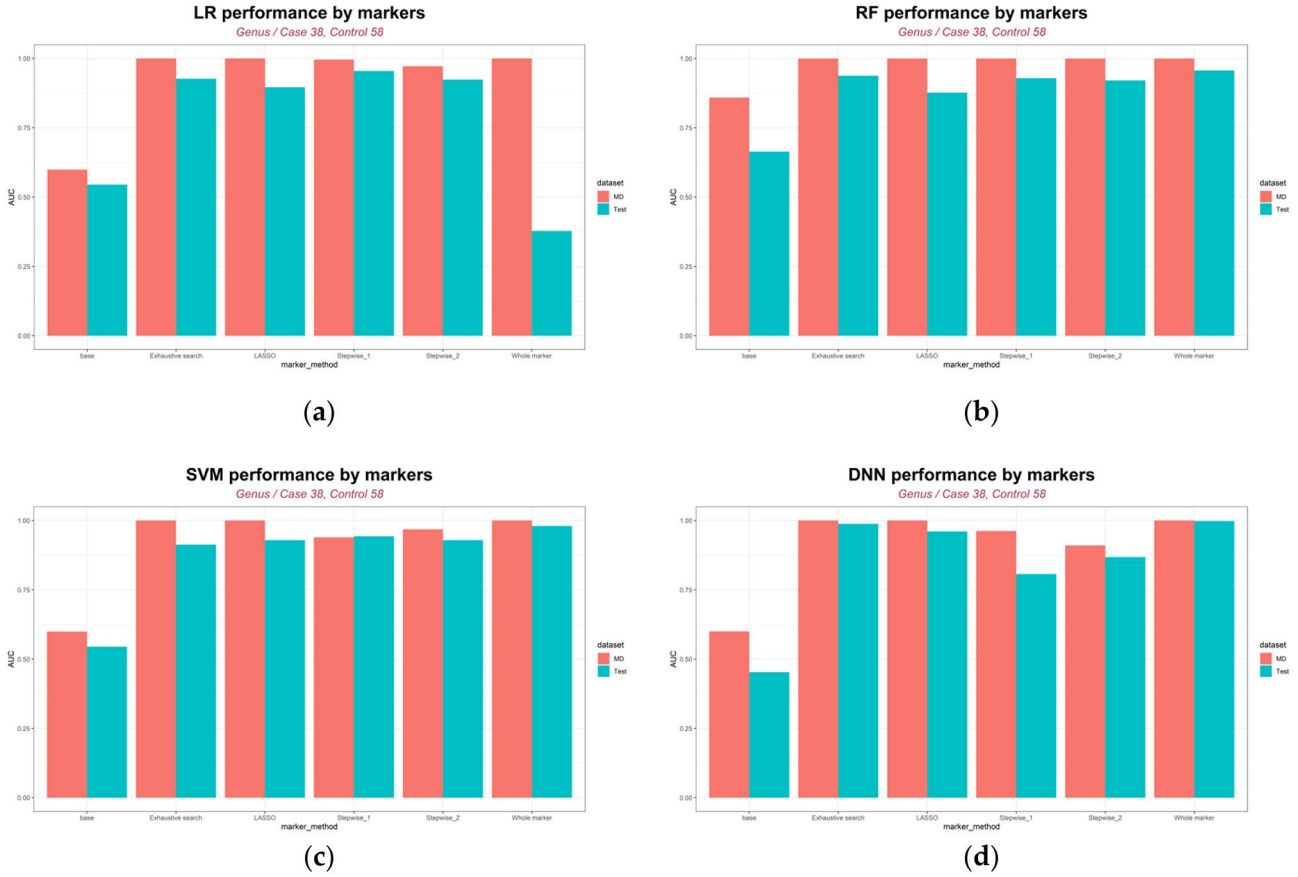


**Fig. 4**. Best model performance by marker set in genus level for test data. (**a**) Performance of LR; (**b**) Performance of RF; (**c**) Performance of SVM; (**d**) Performance of DNN.

LASSO methods achieved 0.937 and 0.899 in test AUC in phylum level, while stepwise2 and stepwise1 in genus level achieved 0.868 and 0.807 test AUC. Here, the full model achieved 0.848 in phylum level and 0.998 in genus level. Overall performance of the model is shown in Figs. 3D and 4D.

For comparison, the best performance for every combination of marker selection method and predictive model is tabulated in Table 4. Compared to the model when only age and sex information were used (null model), the results suggest that the prediction model became more accurate when marker selection was processed at both levels. Also, compared to the result with whole markers without marker selection (full model), the reduced model achieved quite a similar or even better score, with AUC much higher than 0.9, suggesting that marker selection via LASSO and stepwise provides reasonable PC biomarker candidates. In the phylum level, the same markers (*Actinobacteria*, *Verrucomicrobia*) as those identified in the previous[13] were selected when using all multiple marker methods. However, markers selected in genus level were quite different according to the selection methods. Among the four methods in genus level, three selected *Propionibacterium* and *Akkermansia*.

## Discussion

Emerging evidence from microbiome analysis has demonstrated that the human microbiota plays a critical role in the carcinogenesis of various cancers and significantly influences therapeutic responses, including chemotherapy and immunotherapy[42-46]. Several studies have suggested that microbiome signatures can serve as predictive markers for multiple cancers, including colorectal, breast, hepatocellular carcinoma, ovarian, and pancreatic cancer[12,47-50].

Our study is one of the few to explore pancreatic cancer biomarkers through extracellular vesicle (EV)-driven microbiome analysis, a novel approach that offers significant advantages over traditional methods. Unlike fecal or tissue-based microbiome analyses, which provide localized microbial signatures, EV-driven microbiome profiling captures systemic microbial alterations by identifying microbial extracellular vesicles circulating in the bloodstream. This systemic approach enables the detection of microbiome shifts linked to tumor microenvironment changes and immune interactions, thereby potentially improving pancreatic cancer diagnostic sensitivity and specificity.

By integrating EV microbiome profiling with machine learning-based biomarker selection, our study achieved high predictive performance (AUC > 0.95). Using multiple feature selection methods (LASSO, stepwise selection) and classification models (DNN, SVM, RF, LR), we optimized the selection of microbiome markers at both the phylum and genus levels. Our results suggest that EV-based microbial markers may offer diagnostic

| | Phylum | | | | | Genus | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Clinical information | Marker selection | Prediction model | $AUC_{MD}$ | $AUC_{ts}$ | Marker selection | | Prediction model | $AUC_{MD}$ | | $AUC_{ts}$ |
| Age, sex | Clinical information only | LR | 0.599 | 0.545 | | Clinical information only | LR | 0.599 | 0.545 | |
| | | RF | 0.859 | 0.664 | | | RF | 0.859 | 0.664 | |
| | | SVM | 0.599 | 0.545 | | | SVM | 0.599 | 0.545 | |
| | | DNN | 0.600 | 0.453 | | | DNN | 0.600 | 0.453 | |
| | Single (2) | LR | 0.970 | 0.964 | | Single (7) | LR | 1.000 | 0.927 | |
| | | RF | 1.000 | 0.952 | | | RF | 1.000 | 0.938 | |
| | | SVM | 0.984 | 0.957 | | | SVM | 1.000 | 0.913 | |
| | | DNN | 0.910 | 0.954 | | | DNN | 1.000 | 0.988 | |
| | Stepwise1 (3) | LR | 0.990 | 0.934 | | Stepwise1 (3) | LR | 0.996 | 0.955 | |
| | | RF | 1.000 | 0.891 | | | RF | 1.000 | 0.929 | |
| | | SVM | 0.984 | 0.957 | | | SVM | 0.939 | 0.943 | |
| | | DNN | 0.855 | 0.959 | | | DNN | 0.962 | 0.807 | |
| | Stepwise2 (3) | LR | 0.955 | 0.915 | | Stepwise2 (3) | LR | 0.972 | 0.924 | |
| | | RF | 1.000 | 0.935 | | | RF | 1.000 | 0.921 | |
| | | SVM | 0.931 | 0.947 | | | SVM | 0.968 | 0.929 | |
| | | DNN | 0.858 | 0.937 | | | DNN | 0.910 | 0.868 | |
| | LASSO (10) | LR | 0.936 | 0.818 | | LASSO (11) | LR | 1.000 | 0.896 | |
| | | RF | 1.000 | 0.922 | | | RF | 1.000 | 0.877 | |
| | | SVM | 1.000 | 0.844 | | | SVM | 1.000 | 0.929 | |
| | | DNN | 0.966 | 0.899 | | | DNN | 1.000 | 0.961 | |
| | Whole marker (20) | LR | 1.000 | 0.891 | | Whole marker (296) | LR | 1.000 | 0.378 | |
| | | RF | 1.000 | 0.956 | | | RF | 1.000 | 0.957 | |
| | | SVM | 1.000 | 0.889 | | | SVM | 1.000 | 0.980 | |
| | | DNN | 1.000 | 0.848 | | | DNN | 1.000 | 0.998 | |

**Table 4**. Performance of prediction models using various marker sets selected by several methods. The numbers in parentheses indicate the number of markers used in the prediction models.

advantages compared to traditional microbiome profiling methods, which can suffer from sample contamination and limited clinical applicability.

A key strength of our study is the identification of specific microbial taxa linked to pancreatic cancer. Among these, *Akkermansia muciniphila* and *Propionibacterium* have been previously associated with host metabolic and immune responses. *Akkermansia*, a genus within *Verrucomicrobia*, is known for its role in gut barrier integrity and glucose homeostasis[40], and its increased abundance in pancreatic cancer patients suggests a potential metabolic link to tumorigenesis. However, whether *Akkermansia* promotes or inhibits tumor progression in pancreatic cancer remains unclear. On the other hand, *Propionibacterium*, a member of the *Actinobacteria* phylum, has been linked to immune modulation, though the primary butyrate-producing bacteria belong to the *Firmicutes* phylum[38,39]. Future studies are needed to determine whether these microbial taxa actively contribute to pancreatic tumorigenesis or merely reflect disease-associated dysbiosis.

Despite these promising findings, our study has some limitations. First, the relatively small sample size may impact the statistical power of our results. While propensity score matching (PSM) was applied to minimize bias from confounding variables, a larger cohort is necessary for validation. Additionally, although our study identified potential microbial markers, the causal relationship between microbiota and tumor progression remains unclear. Further research, including functional experiments and longitudinal studies, is needed to establish whether these microbial shifts drive tumorigenesis or are secondary effects.

In summary, our study highlights the potential of EV-driven serum microbiome analysis as a promising tool for pancreatic cancer diagnosis. By integrating advanced machine learning approaches with microbiome biomarker discovery, we provide a framework for future research aimed at improving early detection strategies. Larger validation studies and mechanistic investigations will be critical to translating these findings into clinically applicable diagnostic tools for pancreatic cancer patients.

## Conclusions

The study considers correlations among microbiomes to propose the use of some microbiome markers to differentiate PC patients and healthy groups. Specifically, we propose some candidate markers in phylum level (*Verrucomicrobia*, *Actinobacteria* and *Proteobacteria*) and genus level (*Ruminococcaceae UCG-013*, *Ruminiclostridium*, *Propionibacterium*, *Lachnospiraceae NK4A136 group*, *Corynebacterium.1*, *Akkermansia*, *Mucispirillum*, *Pseudomonas*, *Diaphorobacter*, *Clostridium sensu stricto 1*, and *Turicibacter)*. Although the markers resulted in high AUC with DNN in predicting PC patient diagnosis, additional studies using larger

cohorts for validation and further experiments to clarify the biological function of biomarkers in relation to PC are needed. Still, the finding heralds the promise of PC diagnosis prediction using EV-driven microbiome data.

## Data availability statement

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to patients' privacy.

Received: 16 December 2024; Accepted: 12 March 2025
Published online: 31 March 2025

## References

1. Zhang, X. et al. Circulating biomarkers for early diagnosis of pancreatic cancer: facts and hopes. *Am. J. cancer Res.* **8**, 332 (2018).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *Cancer J. Clin.* **68**, 7–30 (2018).
3. Son, B. et al. The role of tumor microenvironment in therapeutic resistance. *Oncotarget* **8**, 3933 (2016).
4. Hosein, A. N., Brekken, R. A. & Maitra, A. Pancreatic cancer stroma: an update on therapeutic targeting strategies. *Nat. Reviews Gastroenterol. Hepatol.* **17**, 487–505 (2020).
5. De La Cruz, M. S. D. & Young, A. P. RUFFIN IV, M. T. Diagnosis and management of pancreatic cancer. *Am. Family Phys.* **89**, 626–632 (2014).
6. He, X. Y. & Yuan, Y. Z. Advances in pancreatic cancer research: moving towards early detection. *World J. Gastroenterology: WJG.* **20**, 11241 (2014).
7. Diaz, S. O. et al. Second trimester maternal urine for the diagnosis of trisomy 21 and prediction of poor pregnancy outcomes. *J. Proteome Res.* **12**, 2946–2957 (2013).
8. Daniels, L. et al. Fecal Microbiome analysis as a diagnostic test for diverticulitis. *Eur. J. Clin. Microbiol. Infect. Dis.* **33**, 1927–1936 (2014).
9. Wang, B., Yao, M., Lv, L., Ling, Z. & Li, L. The human microbiota in health and disease. *Engineering* **3**, 71–82 (2017).
10. Elinav, E., Garrett, W. S., Trinchieri, G. & Wargo, J. The cancer Microbiome. *Nat. Rev. Cancer*. **19**, 371–376 (2019).
11. van Noort, V. et al. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling. *Cancer Res.* **74**, 5690–5699 (2014).
12. Half, E. et al. Fecal Microbiome signatures of pancreatic cancer patients. *Sci. Rep.* **9**, 16801 (2019).
13. Ren, Z. et al. Gut microbial profile analysis by miseq sequencing of pancreatic carcinoma patients in China. *Oncotarget* **8**, 95176 (2017).
14. Chronopoulos, A. & Kalluri, R. Emerging role of bacterial extracellular vesicles in cancer. *Oncogene* **39**, 6951–6960 (2020).
15. Cho, E. J. et al. Urinary microbiome-based metagenomic signature for the noninvasive diagnosis of hepatocellular carcinoma. *Br. J. Cancer*. **130**, 970–975. https://doi.org/10.1038/s41416-024-02582-1 (2024).
16. Yu, W. et al. Exosome-based liquid biopsies in cancer: opportunities and challenges. *Ann. Oncol.* **32**, 466–477 (2021).
17. Noerholm, M. et al. RNA expression patterns in serum microvesicles from patients with glioblastoma multiforme and controls. *BMC cancer*. **12**, 1–11 (2012).
18. Kim, D. J. et al. Colorectal cancer diagnostic model utilizing metagenomic and metabolomic data of stool microbial extracellular vesicles. *Sci. Rep.* **10**, 2860 (2020).
19. Kim, S. I. et al. Metagenomic analysis of serum microbe-derived extracellular vesicles and diagnostic models to differentiate ovarian cancer and benign ovarian tumor. *Cancers* **12**, 1309 (2020).
20. Kim, J. R. et al. Microbiome markers of pancreatic cancer based on bacteria-derived extracellular vesicles acquired from blood samples: a retrospective propensity score matching analysis. *Biology* **10**, 219 (2021).
21. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*. **10**, 1200–1202 (2013).
22. Peng, X., Li, G. & Liu, Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. *J. Comput. Biol.* **23**, 102–110 (2016).
23. Mandal, S. et al. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Disease*. **26**, 27663 (2015).
24. Lee, C., Lee, S. & Park, T. in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 1777–1781 (IEEE). (2017).
25. Wilcoxon, F. Breakthroughs in statistics. *Individual Comparisons Ranking Methods*, 196–202 (1992).
26. Love, M. I., Huber, W. & Anders, S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
27. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
28. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
29. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*. **7**, 335–336 (2010).
30. Haukoos, J. S. & Lewis, R. J. The propensity score. *Jama* **314**, 1637–1638 (2015).
31. Ali, M. S. et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *J. Clin. Epidemiol.* **68**, 122–131 (2015).
32. Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C. & Sölkner, J. Evaluation of the Lasso and the elastic net in genome-wide association studies. *Front. Genet.* **4**, 270 (2013).
33. Kingma, D. P. & Adam A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
34. Caruana, R., Lawrence, S. & Giles, C. Overfitting in neural Nets: backpropagation, conjugate gradient, and early stopping. *Adv. Neural. Inf. Process. Syst.* **13** (2000).
35. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **58**, 267–288 (1996).
36. Efroymson, M. A. Multiple regression analysis. *Mathematical methods for digital computers*, 191–203 (1960).
37. Esquivel-Elizondo, S., Ilhan, Z., Garcia-Peña, E. & Krajmalnik-Brown, R. Insights into butyrate production in a controlled fermentation system via gene predictions. *MSystems* **2**, 101128msystems00051–101128msystems00017 (2017).
38. Inan, M. S. et al. The luminal short-chain fatty acid butyrate modulates NF-κB activity in a human colonic epithelial cell line. *Gastroenterology* **118**, 724–734 (2000).
39. Saus Martínez, E., Guzman, I., Willis, S., Brunet-Vega, J. R. & Gabaldón Estevan, J. A. A. Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. *Mol Aspects Med.* ; 69: 93–106 (2019). (2019).
40. Shin, N. R. et al. An increase in the Akkermansia spp. Population induced by Metformin treatment improves glucose homeostasis in diet-induced obese mice. *Gut* **63**, 727–735 (2014).
41. Miyabayashi, K., Ijichi, H. & Fujishiro, M. The role of the Microbiome in pancreatic cancer. *Cancers* **14**, 4479 (2022).
42. Geller, L. T. et al. Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).
43. Structure, function and diversity of the healthy human microbiome. *nature* **486**, 207–214 (2012).

44. Xavier, J. B. et al. The cancer microbiome: distinguishing direct and indirect effects requires a systemic view. *Trends in cancer* **6**, 192–204 (2020).
45. Zeller, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology* **10**, 766 (2014).
46. Panebianco, C., Andriulli, A. & Pazienza, V. Pharmacomicrobiomics: exploiting the drug-microbiota interactions in anticancer therapies. *Microbiome* **6**, 1–13 (2018).
47. Chen, Y. et al. Characterization of fecal microbial communities in patients with liver cirrhosis. *Hepatology* **54**, 562–572 (2011).
48. Cho, E. J. et al. Circulating microbiota-based metagenomic signature for detection of hepatocellular carcinoma. *Scientific reports* **9**, 7536 (2019).
49. Thomas, A. M. et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature medicine* **25**, 667–678 (2019).
50. Zhu, J. et al. Breast cancer in postmenopausal women is associated with an altered gut metagenome. *Microbiome* **6**, 1–13 (2018).


## Author contributions

D.L. and C.L. contributed equally to this work. Conceptualization, D.L., C.L., T.P. and J.-Y.J.; methodology D.L. and K.H.; formal analysis, D.L, C.L. and K.H.; investigation, C.L., D.L.; resources, J.-Y.J, Y.H., and W.K.; data curation J.-Y.J, Y.H., and W.K.; writing—original draft preparation, C.L., T.P., J.-Y.J, D.L.; writing—review and editing, T.G., B.K., T.P., J.-Y.J, D.L., S.L., C.L., and K.H.; funding acquisition, T.P. and J.-Y.J. All authors have read and agreed to the published version of the manuscript.

## Funding

The work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2022R1A2C1092497) for TP and the Ministry of Science (2022R1F1A1074343) for SL.

## Declarations

## Conflicts of interest

The authors declare no competing interests.

## Institutional review board statement

This study's protocols were approved by the Institutional Review Board at Seoul National University Hospital (IRB No. 1601-137-739).

## Informed consent statement

The informed consent was waived under IRB approval by the Institutional Review Board at Seoul National University Hospital. This study is a retrospective designed study only using data from blood samples already obtained.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94183-y.

**Correspondence** and requests for materials should be addressed to J.-Y.J. or T.P.