Cell Reports Medicine, Volume *4*

Supplemental information

A blood-based metabolomic signature

predictive of risk for pancreatic cancer

Ehsan Irajizad, Ana Kenney, Tiffany Tang, Jody Vykoukal, Ranran Wu, Eunice Murage, Jennifer B. Dennison, Marta Sans, James P. Long, Maureen Loftus, John A. Chabot, Michael D. Kluger, Fay Kastrinos, Lauren Brais, Ana Babic, Kunal Jajoo, Linda S. Lee, Thomas E. Clancy, Kimmie Ng, Andrea Bullock, Jeanine M. Genkinger, Anirban Maitra, Kim-Anh Do, Bin Yu, Brian M. Wolpin, Sam Hanash, and Johannes F. Fahrmann

Contents

Supplementary Table S1 (related to Table 3). Patient characteristics for the PLCO Development Set and the set-aside Test Set.

	Development Set					
	Training Set		**Validation Set**		**Set-Aside Test Set**	
	Non-cases	**Cases**	**Non-cases**	**Cases**	**Non-cases**	**Cases**
Total	494	102	142	33	225	37
Gender, N (%)						
Female	204 (41)	41 (40)	61 (43)	17 (52)	91 (40)	13 (35)
Male	290 (59)	61 (60)	81 (57)	16 (48)	134 (60)	24 (65)
Age At Randomization, N (%)						
<= 59	116 (23)	21 (21)	22 (15)	11 (33)	45 (20)	5 (14)
60-64	108 (22)	24 (24)	34 (24)	3 (9)	63 (28)	14 (38)
65-69	192 (39)	41 (40)	50 (35)	9 (27)	79 (35)	13 (35)
>= 70	78 (16)	16 (16)	36 (25)	10 (30)	38 (17)	5 (14)
Race, N (%)						
White	463 (94)	99 (97)	107 (75)	24 (73)	211 (94)	33 (89)
Black	22 (4)	3 (3)	2 (1)	1 (3)	6 (3)	2 (5)
Other	9 (2)	0 (0)	33 (23)	8 (24)	8 (4)	2 (5)

Supplementary Table S2 (related to Table 2). Selected microbial-associated metabolites and corresponding model coefficients in LASSO regression.

	Lasso selection	
Metabolite	**Selected in model**	**Coefficient**
AcetylCadaverine	-	-
5-hydroxy-L-tryptophan	-	-
5-methoxy-3-indoleacetic acid	-	-
Indole-3-lactic acid	-	-
Indoleacrylic acid	Yes	0.3653
Glycodeoxycholate	-	-
Indole-3-acetaldehyde	-	-
Indole-3-ethanol	-	-
Indole-derivative_2	Yes	0.5022
Indole-derivative_1	-	-
TMAO	Yes	0.2412
Deoxycholate	-	-
Indole-3-acetamide	-	-
Indole-3-acetate	-	-

Supplementary Table S3 (related to Table 2). Stability check of the LASSO regression using perturbed training data and evaluated on the Validation Set for the 3-marker microbial panel.

	Perturbations	AUC (95% CI)	Adj OR†
Lasso regression with 3 selected features	2 randomly selected centers	0.63 (0.44-0.82)	1.37 (0.89-2.09)
	2 randomly selected centers	0.73 (0.60-0.86)	2.33 (1.52-3.77)
	2 randomly selected centers	0.54 (0.41-0.68)	1.25 (0.90-1.73)
	2 randomly selected centers	0.55 (0.45-0.63)	1.27 (0.92-1.72)
	3 randomly selected centers	0.64 (0.54-0.73)	1.65 (1.23-2.24)
	300 random samples	0.60 (0.51-0.68)	1.40 (1.02-1.90)

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs)

Supplementary Table S4 (related to Table 3). Performance of the 3-marker microbial panel, the 5-marker non-microbial panel, and the combined (microbial+non-microbial) metabolite pannel amongst diabetic and non-diabetic individuals in the PLCO set-aside Test Set. † Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase. N0: Number of non-cases, N1: Number of cases.

	3-marker microbial panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.62	0.8	0.77	N0 = 210	0.64	1.84	<0.001
	N1 = 4	(0.22-1.00)	(0.09-3.61)		N1 = 33	(0.53-0.77)	(1.32-2.61)	
All PLCO samples	N0 = 55	0.6	1.56	0.13	N0 = 805	0.62	1.5	<0.001
	N1 = 22	(0.46-0.74)	(0.88-2.95)		N1 = 150	(0.57-0.67)	(1.27-1.77)	
	5-marker non-microbial panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.65	1.93	0.43	N0 = 210	0.75	2.74	<0.001
	N1 = 4	(0.27-1.00)	(0.45-17.61)		N1 = 33	(0.65-0.84)	(1.83-4.32)	
All PLCO samples	N0 = 55	0.67	2.67	0.004	N0 = 805	0.74	2.95	<0.001
	N1 = 22	(0.52-0.82)	(1.44-5.72)		N1 = 150	(0.70-0.78)	(2.12-3.20)	
	Combined (microbial+non-microbial) Panel							
	Diabetics				**Non-Diabetic**			
	Sample Size	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14	0.65	1.7	0.52	N0 = 210	0.81	3.39	<0.001
	N1 = 4	(0.29-1.00)	(0.38-13.34)		N1 = 33	(0.72-0.89)	(2.19-5.61)	
All PLCO samples	N0 = 55	0.67	2.71	0.004	N0 = 805	0.76	2.79	<0.001
	N1 = 22	(0.53-0.81)	(1.44-5.84)		N1 = 150	(0.72-0.80)	(2.27-3.46)	

Supplementary Table S5 (related to Table 3). Patient and tumor characteristics for the newly-diagnosed PDAC cohort.

Variable	PDAC Case (N=99)		Chronic Pancreatitis (N=50)		Healthy Control (N=100)	
	No.	**%**	**No.**	**%**	**No.**	**%**
Institution						
DF/BWCC	69	70%	30	60%	94	94%
BIDMC	15	15%	15	30%	0	0%
CUMC	15	15%	5	10%	6	6%
Age (year), median (IQR)	69.8 (62.5-74.8)		65.4 (54.7-72.2)		63.7 (55.7-70.6)	
Gender						
Male	51	52%	33	66%	51	51%
Female	48	48%	17	34%	49	49%
Race						
White	94	95%	42	84%	84	86%
Black/African-American	0	0%	5	10%	5	5%
Asian	1	1%	0	0%	2	2%
Other	4	4%	3	6%	7	7%
Blood collection year						
2015-2016	19	19%	2	4%	0	0%
2017-2019	80	81%	48	96%	100	100%
Smoking Status						
Current Smoker	6	6%	11	22%	4	4%
Past smoker	50	51%	17	34%	42	42%
Never smoker	43	43%	22	44%	54	54%
BMI (kg/m^2), Meidan(IQR)	27.4 (24.0-30.0)		25.0 (22.8-27.6)		27.5 (24.3-32.0)	
Diabetes						
No	64	65%	23	46%	93	93%
Yes	35	35%	27	54%	7	7%
Etiology of chronic pancreatitis						

Alcohol	-	-	16	32%	-	-
Autoimmune	-	-	2	4%	-	-
Congenital anatomical variant	-	-	3	6%	-	-
Duct stricture or stones	-	-	7	14%	-	-
Idiopathic	-	-	21	42%	-	-
Other	-	-	1	2%	-	-
AJCC 8th edition staging pTNM[a]						
T0-2N0M0	15	24%	-	-	-	-
T3-4N0M0	2	3%	-	-	-	-
T1-4N1M0	28	45%	-	-	-	-
T1-4N2M0	17	28%	-	-	-	-
AJCC 8th edition staging ypTNM[b]						
T0-2N0M0	24	64%	-	-	-	-
T3-4N0M0	1	3%	-	-	-	-
T1-4N1M0	7	19%	-	-	-	-
T1-4N2M0	5	14%	-	-	-	-
PDAC recurrence						
No[c]	56	57%	-	-	-	-
Yes	43	43%	-	-	-	-

DF/BWCC: Dana-Farber/Brigham and Women's Cancer Center; BIDMC: Beth Israel Deaconess Medical Center; CUMC: Columbia University Medical Center

AJCC: American Joint Committee on Cancer, PDAC: Pancreatic ductal adenocarcinoma, BMI: Body mass index

[a] Patients who underwent up-front surgical resection

[b] Patients who received neoadjuvant treatment and then underwent surgical resection

[c] The median (IQR) follow-up time was 15.0 (7.2-23.2) months for patients without cancer recurrence

Supplementary Table S6 (related to Figure 2). Performance of all non-microbial metabolites in the PLCO Training and Validation Sets.

See excel file.

Supplementary Table S7 (related to Table 3). Selected non-microbial metabolites.

	Training - 5 centers		Validation- 2 centers	
Name	**Adj. Odds Ratio†**	***P-value (FDR) ‡***	**Adj. Odds Ratio†**	***P-value£***
Cholesterol glucuronide	1.735	<0.001	1.720	0.006
Galactosamine	1.749	<0.001	1.514	0.035
2-Hydroxyglutarate	1.857	<0.001	1.738	0.006
Erythritol	1.688	<0.001	1.532	0.030
Glucose	1.744	<0.001	1.662	0.018

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase

‡ Benjamini and Hochberg adjusted p-values

£ Raw p-values

Supplementary Table S8 (related to Table 3). Performance of different learning models based on non-microbial metabolites and model stability check in the PLCO Validation Set. † Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase.

Model	Hyperparameters	AUC (95% CI)	Adj OR†
Logistic regression	-	0.72 (0.63-0.81)	2.10 (1.04-2.90)
logistic regression with ridge (L2) regularization	Penalty weight = 0.18	0.69 (0.58-0.78)	1.74 (1.20-2.25)
logistic regression with LASSO (L1)	Penalty weight = 0.01, number of selected features = 4	0.71 (0.54-0.73)	2.08 (0.94-2.83)
Iterative Random Forest	Number of iterations = 3	0.60 (0.49-0.72)	1.44 (0.90-1.90)
Deep neural network model	Number of cross-validation folds = 6, hidden layers = 3 with 32 nodes in each layer	0.59 (0.48-0.68)	1.43 (0.95-2.10)
GBM	Number of trees = 42, max depth= 5	0.58 (0.46-0.67)	1.30 (0.93-1.87)
Auto ML	Selected model = randomized trees	0.66 (0.52-0.72)	1.85 (1.50-2.02)

	Perturbations	AUC (95% CI)	Adj OR†
Logistic regression with 5 selected features	2 randomly selected centers	0.71 (0.52-0.87)	2.10 (1.10-2.94)
	2 randomly selected centers	0.74 (0.61-0.91)	2.33 (1.42-4.10)
	2 randomly selected centers	0.69 (0.59-0.80)	2.11 (0.90-2.73)
	2 randomly selected centers	0.67 (0.45-0.85)	1.90 (1.12-2.72)
	3 randomly selected centers	0.60 (0.52-0.68)	1.65 (1.23-2.24)
	300 random samples	0.64 (0.55-0.71)	1.73 (1.52-2.20)

Supplementary Table S9 (related to Table 3). Performance of the 5-marker non-microbial panel in the PLCO set-aside Test Set and the entire specimen set.

Set-aside Test Set				
		5-marker non-microbial panel [a]		
Time to Dx	Sample Size	AUC. (95% CI)	Adj. OR **†** (95% CI)	*P-value*

[0-5)	N0 = 225
	N1 = 37
		0.74
		(0.65 - 0.83)
			2.72
			(1.83 - 4.24)
				<0.001

[0-2)	N0 = 225
	N1 = 24
		0.82
		(0.72 - 0.92)
			4.03
			(2.41 - 7.32)
				<0.001

[2-5)	N0 = 225
	N1 = 13
		0.59
		(0.44 - 0.72)
			1.32
			(0.71 - 2.41)
				0.36

Entire Set (Development + Set-aside Test Set)				
		5-marker non-microbial panel [a]		
Time to Dx	Sample Size	AUC. (95% CI)	Adj. OR **†** (95% CI)	*P-value*

[0-5)	N0 = 861
	N1 = 172
		0.74
		(0.67 - 0.77)
			2.59
			(2.13 - 3.18)
				<0.001

[0-2)	N0 = 861
	N1 = 92
		0.80
		(0.75 - 0.85)
			3.69
			(2.83 - 4.91)
				<0.001

[2-5)	N0 = 861
	N1 = 80
		0.65
		(0.59 - 0.72)
			1.74
			(1.37 – 2.21)
				<0.001

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase

N0: number of non-cases

N1: number of cases

a: Non-microbial-related metabolite signature includes cholesterol glucuronide, hydroxyglutarate, galactosamine, glucose, and erythritol

Supplementary Table S10 (related to Figure 2 and Table 4). Performance of the combined metabolite panel plus CA19-9 stratified by diabetic status.

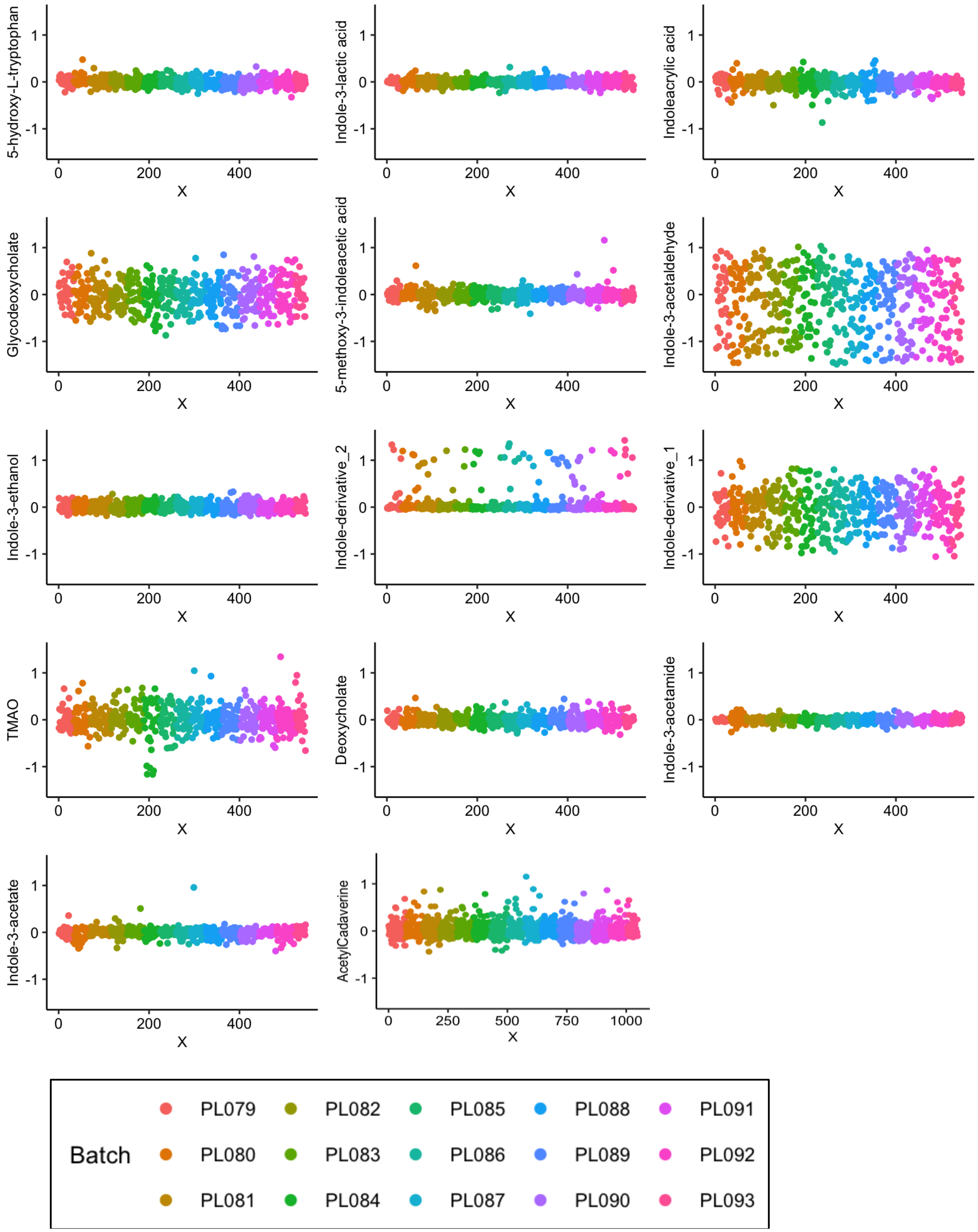
	Diabetics				**Non-Diabetic**			
5-marker non-microbial panel + 3-marker microbial panel + CA19.9	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI) †**	**P-value**	**Sample Size**	**AUC (95% CI)**	**Adj. OR (95% CI)†**	**P-value**
PLCO Testing Set	N0 = 14 N1 = 4	0.78 (0.50-1.00)	6.82 (1.14-210.61)	0.10	N0 = 210 N1 = 33	0.84 (0.76-0.92)	10.21 (4.55-26.61)	<0.001
All PLCO samples	N0 = 55 N1 = 22	0.71 (0.60-0.84)	3.75 (1.81-9.72)	0.001	N0 = 805 N1 = 150	0.80 (0.76-0.84)	9.54 (6.36-14.75)	<0.001

† Age, gender, BMI, and smoking status were included as covariates in adjusted odds ratios (ORs); odds ratio per unit SD increase

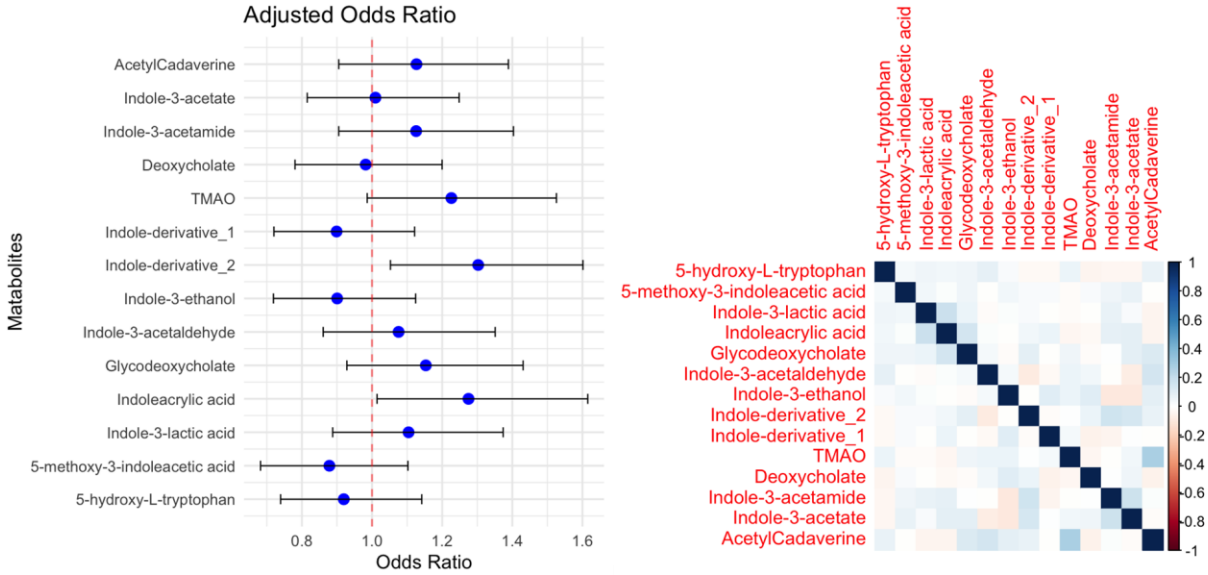
N0: number of non-cases

N1: number of cases

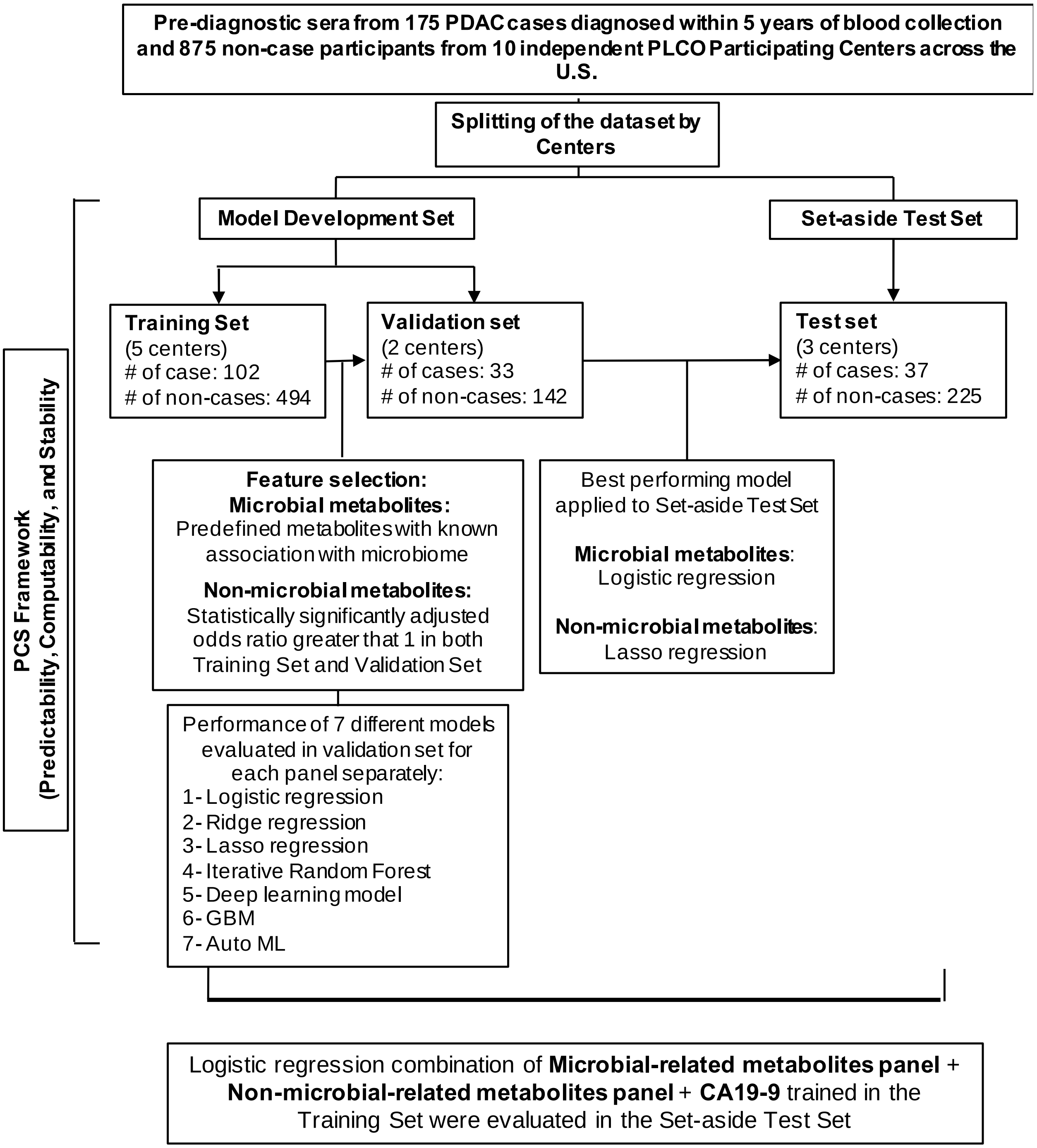
Supplementary Figure S1 (related to Figure 1 and Table 2). Distribution plots for detected microbial-related metabolites across analytical batches in the PLCO specimen set. X-axis represents individual specimens.



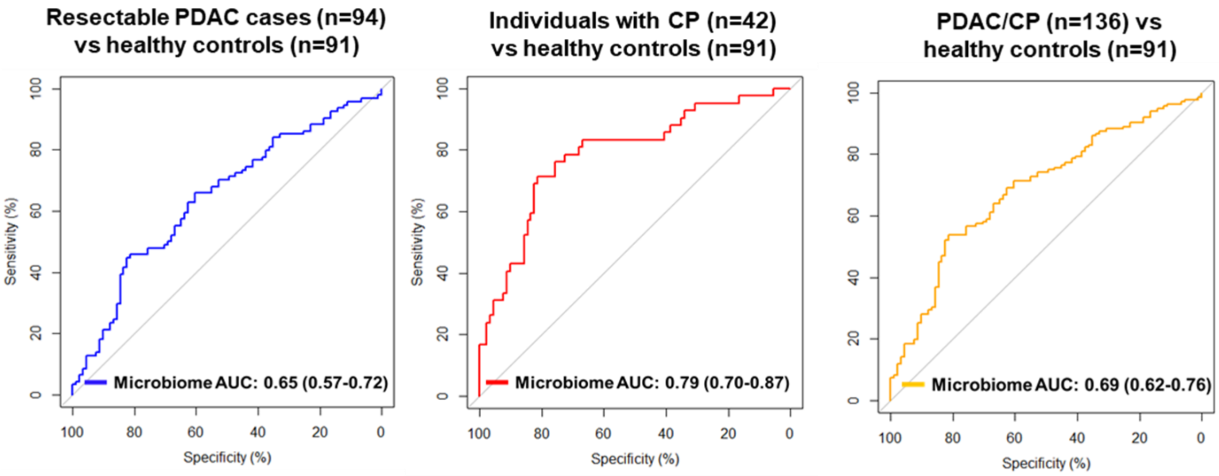
Supplementary Figure S2 (related to Figure 1 and Table 2). Odds ratios, adjusted odds ratios, and correlations for individual microbial-related metabolites for risk of pancreatic cancer in the Training Set. Gender, age, smoking status, and BMI were included as covariates in adjusted odds ratios.



Supplementary Figure S3 (related to Table 2). Workflow of analyses.


Pre-diagnostic sera from 175 PDAC cases diagnosed within 5 years of blood collection and 875 non-case participants from 10 independent PLCO Participating Centers across the U.S.
Splitting of the dataset by Centers
Model Development Set
Set-aside Test Set
Training Set
(5 centers)
# of case: 102
# of non-cases: 494
Validation set
(2 centers)
# of cases: 33
# of non-cases: 142
Test set
(3 centers)
# of cases: 37
# of non-cases: 225
PCS Framework
(Predictability, Computability, and Stability
Feature selection:
Microbial metabolites:
Predefined metabolites with known association with microbiome
Non-microbial metabolites:
Statistically significantly adjusted odds ratio greater that 1 in both Training Set and Validation Set
Best performing model applied to Set-aside Test Set
Microbial metabolites: Logistic regression
Non-microbial metabolites: Lasso regression
Performance of 7 different models evaluated in validation set for each panel separately:
1- Logistic regression
2- Ridge regression
3- Lasso regression
4- Iterative Random Forest
5- Deep learning model
6- GBM
7- Auto ML
Logistic regression combination of Microbial-related metabolites panel + Non-microbial-related metabolites panel + CA19-9 trained in the Training Set were evaluated in the Set-aside Test Set

Supplementary Figure S4 (related to Table 3). Predictive performance of the 3-marker microbial panel in the independent newly-diagnosed PDAC cohort. Abbreviation: CP- chronic pancreatitis. A subset samples were excluded due to insufficient sample volume or not having passed quality control criteria.



Resectable PDAC cases (n=94) vs healthy controls (n=91)

Individuals with CP (n=42) vs healthy controls (n=91)

PDAC/CP (n=136) vs healthy controls (n=91)

Odds Ratio (95% CI)		
Resectable PDAC cases vs healthy controls	Individuals with CP vs healthy controls	PDAC/CP vs healthy controls
1.55 (1.13-2.23)	2.83 (1.83-4.82)	2.07 (1.45-3.18)