Cell Reports Medicine, Volume *5*

Supplemental information

Circulating microbiome DNA as biomarkers for early diagnosis and recurrence of lung cancer

Haiming Chen, Yi Ma, Juqing Xu, Wenxiang Wang, Hao Lu, Cheng Quan, Fan Yang, Yiming Lu, Hao Wu, and Mantang Qiu

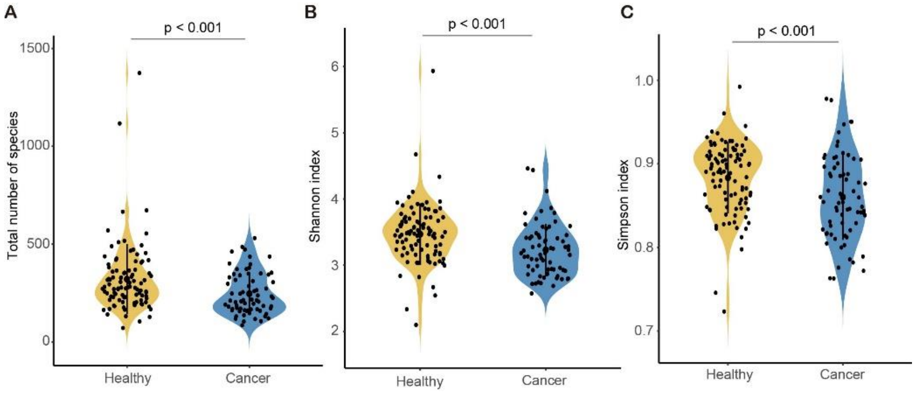

Figure S1 Alpha diversity between cancer patients and healthy controls in the training cohort. Related to Figures 1.

(A-C) (A) Total number of detected species, (B) Shannon diversity index and (C) Simpson diversity index was computed from all samples in the healthy and cancer groups.

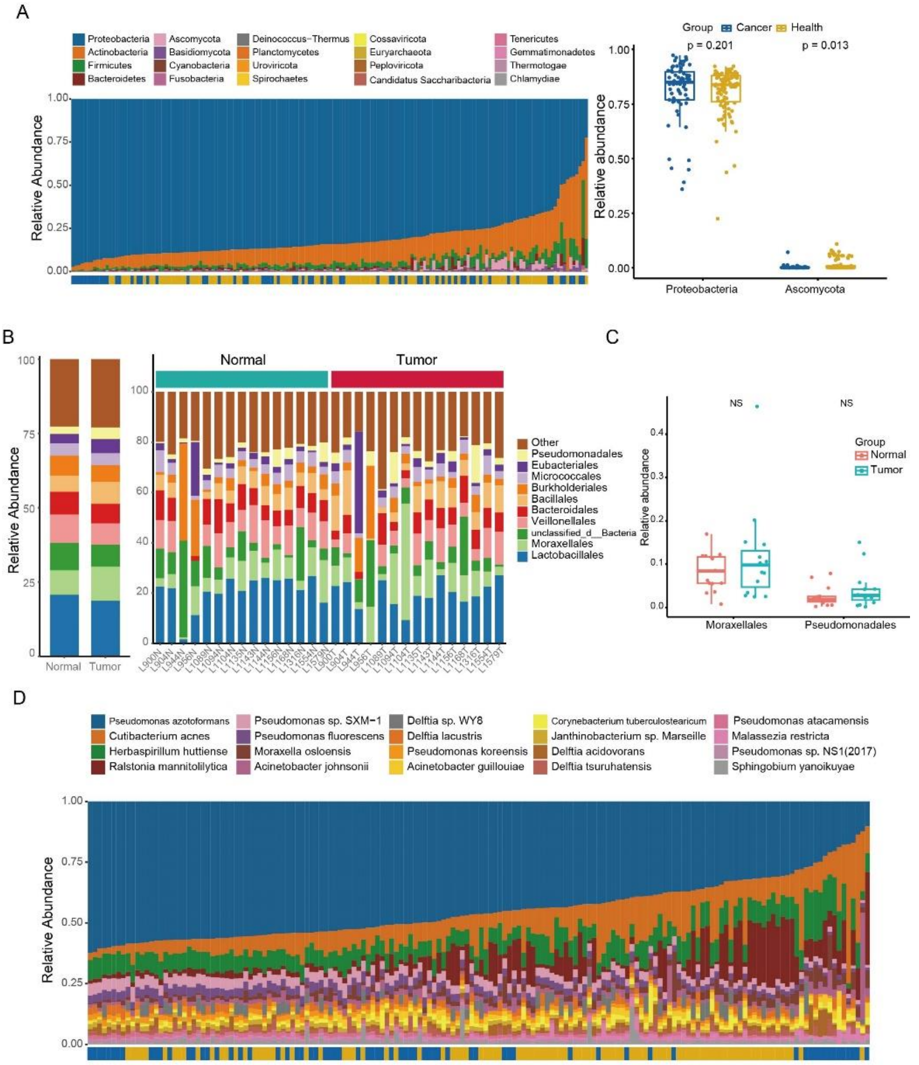

Figure S2 Tumor microbiome composition in the training dataset from cancer Detection cohort. Related to Figures 2.

(A) Circulating microbiome composition at the phylum level in the training dataset of Detection model (ordered by the most abundant taxa, *Proteobacteria* phylum). (B) Intratumour microbiome composition at order taxonomic levels in tumor tissue and paired normal tissue. (C) Boxplots showed relative abundance of specific order-level taxa in the intratumour microbiome analysis. (D) circulating microbiome composition at the species level in the training dataset of Detection model (ordered by the most abundant taxa, *Pseudomonas azotoformans* species).

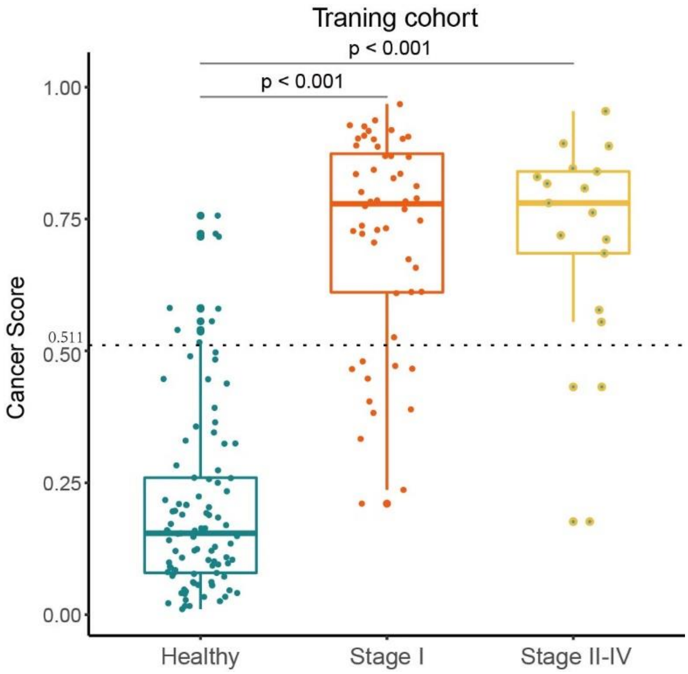

Figure S3 The boxplots showing the distribution of cancer scores of healthy controls, and cancer patients with different TNM stages in the training cohort. Related to Figures 3.

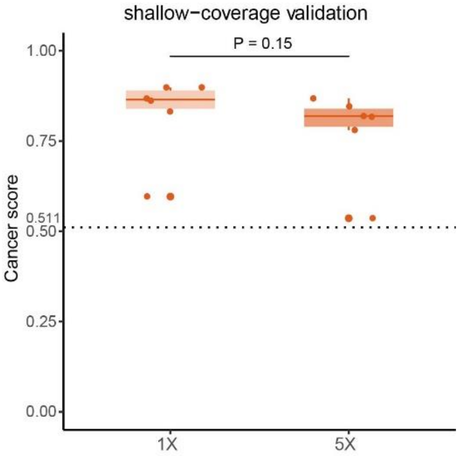

Figure S4 The boxplots showing the distribution of cancer scores of the additional shallow-coverage dataset in coverage depths of 1X and 5X, and a Wilcoxon test was performed for the comparison between the 1X and the 5X. Related to Figures 4.

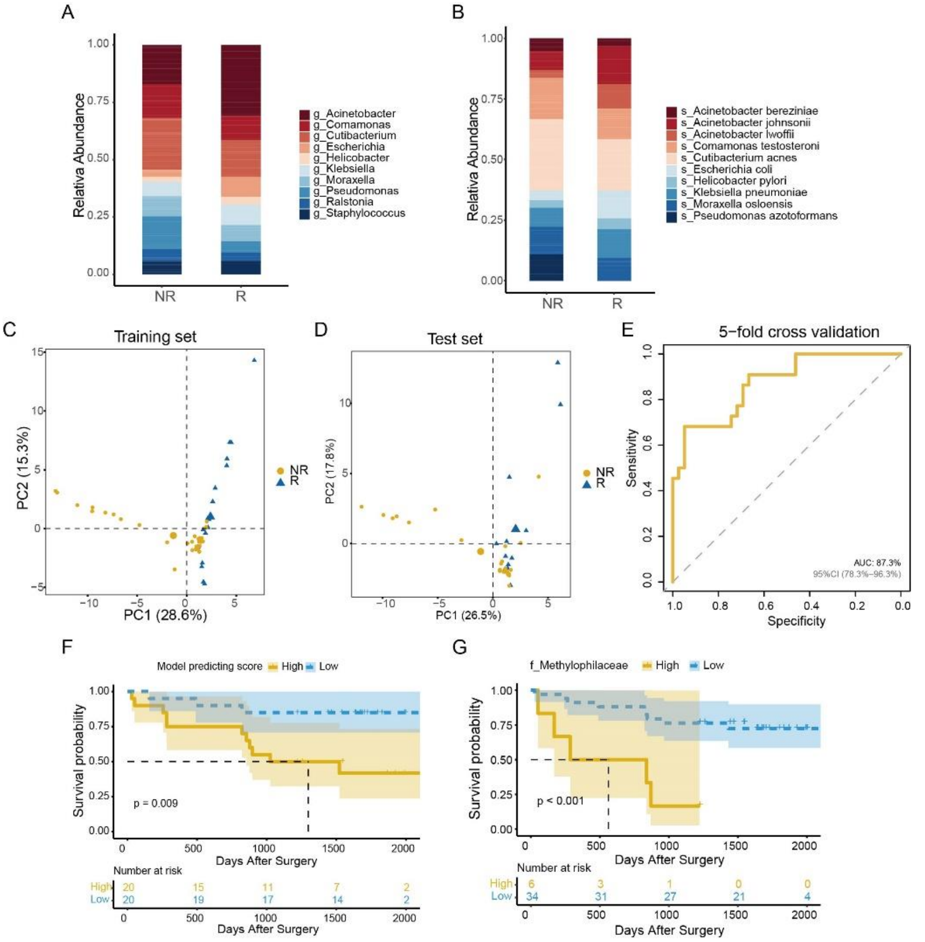

Figure S5 Differentially enriched circulating microbial taxa are associated with RFS of lung cancer. Related to Figures 5 and 6.

(A-B) The Bar plots of the genera (A) and species (B) taxonomic levels in recurrence and non-recurrence lung cancer patients in the training set. (C-D) Principle component analysis showed stratification of samples in the training set (C) and the test set by significant taxa relative abundance, respectively. PC1 and PC2 values represented the top two principal coordinates. Different sample types were denoted by color code and shape. (E) Receiver operating characteristic curve of 5-fold cross-validation in the training set. (F) Kaplan-Meier plot of lung cancer patients defined by recurrence model predicting scores in the test set. (G) Kaplan-Meier estimates for RFS probability of patients with different abundance levels of *Methylophilaceae* family.

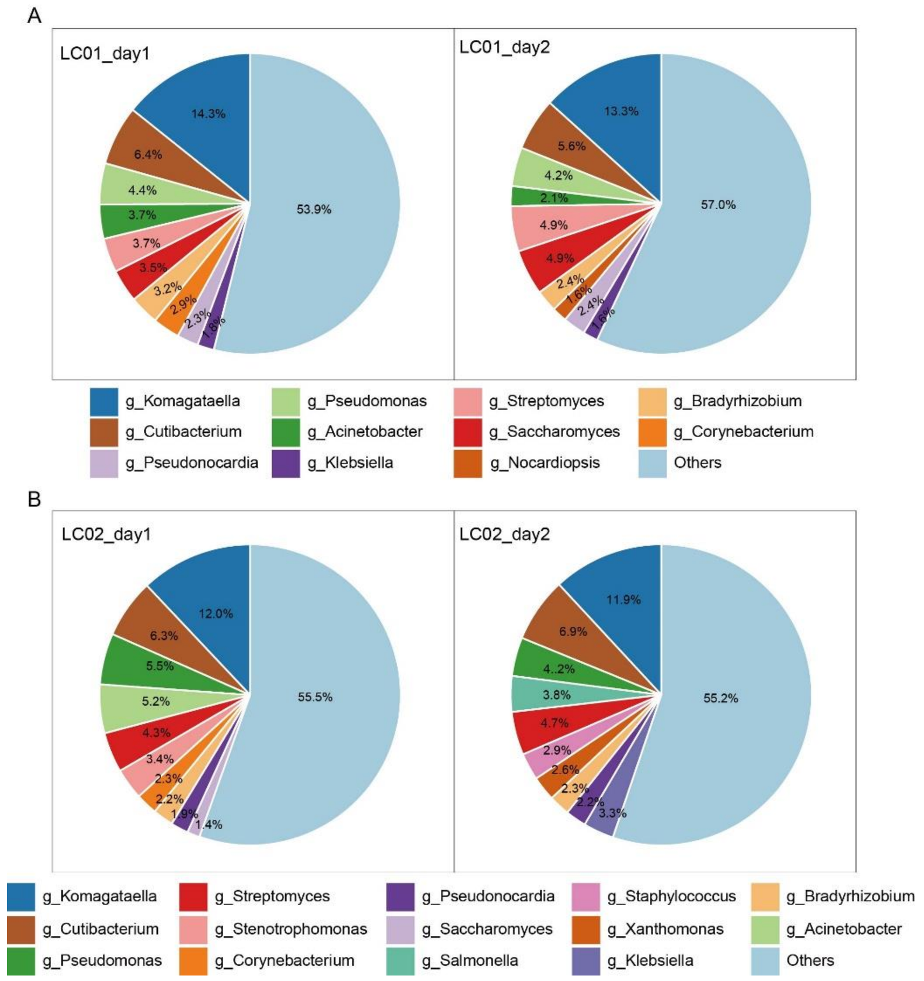

Figure S6 The circulating microbial DNA profiles both the day before and the day of surgery. Related to the STAR Methods.

Pie charts showed the most abundance genus in the patient LC01 (A) and the patient LC02 (B) both the day before and the day of surgery.

Table S3. Clinical characteristics of study participants in the training and validation cohorts and their disposition, related to Figure 1.

Cohort	Training (N = 166)			Validation I (N = 96)			Validation II (N = 53)		
	Cancer (n = 69)	Healthy (n = 97)	p	Cancer (n = 48)	Healthy (n = 48)	p	Cancer (n = 33)	Healthy (n = 20)	p
Age in years, Mean±SD	58.16 ±9.99	59.31 ±13.23	0.526	59.79 ±8.207	60.08 ±10.946	0.148	69.09 ±12.428	66.15 ±16.731	0.501
Cigarette smoking, n (%)	25 (36.2%)	37 (38.1%)	0.802	7 (14.6%)	8 (16.7%)	0.779	5 (15.2%)	1 (5.0%)	0.258
Female, *n* (%)	31 (44.9%)	32 (33.0%)	0.118	27 (56.3%)	28(58.3 %)	0.837	18 (54.5)	15 (75.0%)	0.136
Histology, *n* (%)									
LUAD	51 (73.9%)	-	-	45 (93.7%)	-		24 (72.7%)	-	-
LUSC	12 (17.4%)	-	-	0 (0.0%)	-		4 (12.1%)	-	-
SCLC	5 (7.2%)	-	-	1 (2.1%)	-		2 (6.1%)	-	-
Others*	1 (1.4%)	-	-	2 (4.2%)	-		3 (9.1%)	-	-
Stage, *n* (%)									
I	52 (75.4%)	-	-	41 (85.4%)	-		0 (0.0%)	-	-
II+III+IV	17(24.6%)	-	-	7 (14.6%)	-		33(100.0%)	-	-
Tumor size, *n* (%)									
< 1 cm	28 (40.6%)	-	-	29 (60.4%)	-		0 (0.0%)	-	-
≥ 1cm	41 (59.4%)	-	-	19 (39.6%)	-		33 (100.0%)	-	-

**Others*: lung adenosquamous carcinoma, large-cell carcinoma of the lung

Table S5. Lung cancer samples enrolled in intratumor microbiome analysis, related to Figure S2.

Sample
L900
L904
L944
L956
L1089
L1094
L1104
L1135
L1143
L1144
L1156
L1168
L1316
L1554
L1579

Table S8. The diagnostic performance of the predictive model in the training cohort, related to Figure 3.

Training cohort		Actual	
		Cancer	**Control**
Predicted	**Cancer**	56	9
	Control	13	88
Sensitivity (95% CI)		81.2% (69.6%-89.2%)	
Specificity (95% CI)		90.7% (82.7%-95.4%	
Accuracy (95% CI)		86.8% (80.6%-91.5%)	

Definition of abbreviation: CI= confidence interval.

Table S10. The diagnostic performance of the predictive model in the validation cohorts, related to Figure 4.

Validation cohort I		**Actual**	
		Cancer	**Control**
Predicted	**Cancer**	42	12
	Control	6	36
Sensitivity (95% CI)		87.5% (74.1%-94.8%)	
Specificity (95% CI)		75.0% (60.1%-85.9%	
Accuracy (95% CI)		81.3% (72.2%-88.5%)	
Validation cohort II		**Actual**	
		Cancer	**Control**
Predicted	**Cancer**	29	2
	Control	4	18
Sensitivity (95% CI)		87.9% (70.9%-96.0%)	
Specificity (95% CI)		90.0% (66.9%-98.2%)	
Accuracy (95% CI)		88.7% (77.0%-95.7%)	
Combined validation cohorts		**Actual**	
		Cancer	**Control**
Predicted	**Cancer**	71	14
	Control	10	54
Sensitivity (95% CI)		87.7% (78.0%-93.6%)	
Specificity (95% CI)		79.4% (67.5%-87.9%)	
Accuracy (95% CI)		83.9% (77.0%-89.4%)	

Definition of abbreviation: CI= confidence interval.

Table S12. The cancer scores of additional shallow-coverage dataset, related to Figure 4.

Sample	**Low coverages**	
	5X	1X
NJ_C0048	0.868	0.898
NJ_C0067	0.536	0.596
NJ_C0070	0.846	0.868
NJ_C0074	0.78	0.832
NJ_C0086	0.82	0.862
NJ_C0092	0.818	0.898

Table S13.Characteristics of the included participants of the recurrence cohort, related to Figure 5.

Characteristics	Train set (N = 61)			Test set (N = 40)		
	R (n = 22)	NR (n = 39)	p	R (n = 14)	NR (n = 26)	p
Recurrence free survival (years), Mean ± SD	1.70±0.87	4.39±0.86	<0.001	1.65±1.15	4.66±0.79	<0.001
Age (years), Mean±SD	61.36±7.14	61.59±8.06	0.913	58.57±8.36	61.81±8.99	0.273
Female, *n* (%)	12 (54.5%)	19 (48.7%)	0.662	8 (57.1%)	10 (38.5%)	0.257
Cigarette smoking, n (%)	10 (45.5%)	11 (28.2%)	0.173	6 (42.9%)	8 (30.8%)	0.445
Stage, n (%)			0.603			0.002
I	13 (59.1%)	27 (69.2%)		5 (35.7%)	22 (84.6%)	
II-III	9 (40.9%)	12 (30.8%)		9 (64.3%)	4 (15.4%)	
Histology, n (%)			0.37			0.232
LUAD	20 (90.9%)	36 (92.3%)		12 (85.7%)	25 (96.2%)	
LUSC	1 (4.5%)	3 (7.7%)		2 (14.3%)	1 (3.8%)	
Others	1 (4.5%)	0 (0.0%)		0 (0.0%)	0 (0.0%)	
Tumor diameters (cm), Mean ± SD	2.29±0.60	2.05±0.64	0.16	1.99±0.72	1.90±0.73	0.725

R, recurrence group; NR, non-recurrence group; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; Others, lung adenosquamous carcinoma

Table S14: Significantly differential taxa between R and NR groups, related to Figure 5.

feature	score	group	LDA	p
s_Limnohabitanssp_103DPR2	3.490936	R	3.092666	0.001628
s_Phreatobactercathodiphilus	1.996731	NR	2.35054	0.036657
s_CandidatusMethylopumilusuniversalis	3.260202	R	2.989751	0.000272
s_Paracoccussanguinis	3.049735	NR	2.579703	0.047382
s_Pseudomonaskoreensis	2.60036	NR	2.339833	0.015943
f_Phreatobacteraceae	1.996731	NR	2.420925	0.036657
g_Paracoccus	3.309408	NR	2.696325	0.03781
o_Nitrosomonadales	3.292351	R	3.022054	8.98E-05
f_Weeksellaceae	3.137434	NR	2.71627	0.017542
o_CandidatusNanopelagicales	3.783817	R	3.363307	0.015678
o_Rhodobacterales	3.312299	NR	2.708654	0.03781
s_CandidatusNanopelagicusabundans	2.946196	R	2.728211	0.009715
g_Malassezia	3.196425	NR	2.869928	0.016941
s_Tepidimonastaiwanensis	2.670287	NR	2.425536	0.024261
g_Tepidimonas	2.670287	NR	2.405391	0.024261
s_Acinetobacterpseudolwoffii	2.10633	R	2.363092	0.019045
g_Psychrobacter	2.443738	NR	2.284315	0.015943
f_Malasseziaceae	3.196425	NR	2.810634	0.016941
s_Pseudomonassp_B10	2.622441	NR	2.319814	0.015943
s_Variovoraxparadoxus	2.368597	NR	2.206128	0.015943
p_Firmicutes	4.078506	NR	3.3144	0.042595
s_Acinetobactersp_NEB149	2.556007	R	2.428587	0.019045
s_Flavobacteriumsp_GENT5	2.592916	R	2.379265	0.019045
g_Sphingomonas	2.565695	NR	2.317423	0.015943
g_Staphylococcus	3.847322	R	3.369032	0.039949
s_Pseudomonaslactis	2.054425	NR	2.463025	0.036657
f_Methylophilaceae	3.292351	R	3.01553	8.98E-05
s_Herbaspirillumhuttiense	2.862965	NR	2.590948	0.015943
s_Sphingomonaspaucimobilis	2.48324	NR	2.347997	0.015943
s_Pseudomonasmoraviensis	2.558204	NR	2.279922	0.036657
g_Limnohabitans	3.603065	R	3.23221	0.001628
s_Acinetobactercalcoaceticus	2.175639	R	2.342373	0.019045
s_Corynebacteriumureicelerivorans	2.116535	NR	2.180723	0.036657
g_Phreatobacter	1.996731	NR	2.285086	0.036657
s_CandidatusPlanktophilavernalis	3.653169	R	3.171637	0.027129
s_Acinetobacteroleivorans	2.394791	R	2.273913	0.006332
s_Psychrobactersanguinis	2.419681	NR	2.250923	0.015943
s_Acinetobacterradioresistens	2.523844	R	2.238038	0.037195
g_CandidatusNanopelagicus	3.105997	R	2.845843	0.003096
o_Propionibacteriales	4.485492	NR	3.775424	0.041089
s_Brevundimonasmediterranea	2.631862	R	2.330522	0.037195

g_CandidatusMethylopumilus	3.274036	R	3.013744	0.000272
s_Pseudomonasazotoformans	4.001611	NR	3.686345	0.006707
s_Pseudomonassp_SXM_1	2.81565	NR	2.53472	0.015943
s_Limnohabitanssp_63ED37_2	2.935637	R	2.725227	0.003097
o_Malasseziales	3.196425	NR	2.848594	0.016941
f_CandidatusNanopelagicaceae	3.783817	R	3.376195	0.015678
s_CandidatusNanopelagicuslimnes	2.572038	R	2.549058	0.002083
g_Variovorax	2.616205	NR	2.360406	0.015943
s_Pseudomonaspoae	2.123864	NR	2.262788	0.036657
s_Paracoccusmarcusii	2.215384	NR	2.076205	0.036657
s_Malasseziarestricta	3.196425	NR	2.799819	0.016941
s_Ralstoniapickettii	2.312351	NR	2.303224	0.036657
s_Pseudomonassp_NS1_2017_	2.2469	NR	2.244133	0.024261
s_Chryseobacteriumsp_ZHDP1	2.868544	NR	2.590359	0.015943
f_Rhodobacteraceae	3.312299	NR	2.69072	0.03781
s_Ralstoniamannitolilytica	3.634975	NR	3.336194	0.015943
s_Acinetobacterpittii	2.395057	R	2.322208	0.019045
g_CandidatusPlanktophila	3.677635	R	3.216565	0.024997
p_Basidiomycota	3.196425	NR	2.845506	0.016941
c_Malasseziomycetes	3.196425	NR	2.854624	0.016941
g_Herbaspirillum	2.862965	NR	2.582135	0.015943

Table S15. Model predicting score adjusted by TNM stage, related to Figure S5.

	Univariable		Multivariable	
	HR (95%CI)	**p**	**HR (95%CI)**	**p**
Predicting score	13.091 (2.236-76.652)	0.004	27.848 (3.581-216.534)	0.001
TNM Stage (II-III vs. I)	5.751 (1.910-17.314)	0.002	7.619 (2.358-24.620)	< 0.001

Table S16. The top 20 of genus-level taxa of circulating microbial DNA profiles both the day before and the day of surgery, related to the STAR Methods.

LC01_day1	LC01_day2	LC02_day1	LC02_day2
g_Komagataella	g_Komagataella	g_Komagataella	g_Komagataella
g_Cutibacterium	g_Cutibacterium	g_Cutibacterium	g_Cutibacterium
g_Pseudomonas	g_Saccharomyces	g_Pseudomonas	g_Streptomyces
g_Acinetobacter	g_Streptomyces	g_Acinetobacter	g_Pseudomonas
g_Streptomyces	g_Pseudomonas	g_Streptomyces	g_Salmonella
g_Saccharomyces	g_Pseudonocardia	g_Stenotrophomonas	g_Klebsiella
g_Bradyrhizobium	g_Bradyrhizobium	g_Corynebacterium	g_Staphylococcus
g_Corynebacterium	g_Acinetobacter	g_Bradyrhizobium	g_Xanthomonas
g_Pseudonocardia	g_Nocardiopsis	g_Pseudonocardia	g_Bradyrhizobium
g_Klebsiella	g_Klebsiella	g_Saccharomyces	g_Pseudonocardia
g_Nocardiopsis	g_Mesorhizobium	g_Mesorhizobium	g_Acinetobacter
g_Malassezia	g_Corynebacterium	g_Nocardiopsis	g_Corynebacterium
g_Staphylococcus	g_Nocardioides	g_Sphingomonas	g_Nocardiopsis
g_Paracoccus	g_Halomonas	g_Brevundimonas	g_Micrococcus
g_Escherichia	g_Paracoccus	g_Nocardioides	g_Burkholderia
g_Actinomyces	g_Microbacterium	g_Microbacterium	g_Nocardioides
g_Brevundimonas	g_Sphingomonas	g_Paracoccus	g_Mesorhizobium
g_Methylobacterium	g_Malassezia	g_Methylobacterium	g_Malassezia
g_Nocardioides	g_Methylobacterium	g_Xanthomonas	g_Sphingomonas
g_Stenotrophomonas	g_Roseomonas	g_Staphylococcus	g_Roseomonas