Gut microbiota as non-invasive diagnostic and prognostic biomarkers for natural killer/T-cell lymphoma

We read with interest the study by Kartal *et al*[1] showing that the gut-microbiota-derived biomarkers for disease stratification are often shared by subjects across disease cohorts. Here, we confirmed their observations with findings from a newly diagnosed natural killer/T-cell lymphoma (NKTCL) cohort, in which the gut biomarkers were significantly overlapped with those of multiple disease cohorts and consistently enriched/depleted in subjects with those diseases. Importantly, many of the shared biomarkers were remarkably associated with patient outcomes in our cohort, implying that they may have broad prognostic values in multiple diseases.

'Microbiota-gut-lymphoma axis' represents a fascinating avenue of microbiota-mediated lymphomagenesis and intervention opportunity,[2] but the implications of gut microbiota in NKTCL remain enigmatic. To identify gut microbiota-derived diagnostic biomarkers for NKTCL, we recruited a discovery cohort consisting of 30 treatment-naïve patients and 20 healthy controls (HCs), and a validation cohort, including 12 patients and 13 HCs, respectively (online supplemental materials and methods). We applied shotgun metagenomic sequencing to their faecal samples, profiled their gut metagenomes using mOTUs2 V.2.5,[3] and trained a patient-stratification classifier with all species-level taxonomic features using the LASSO algorithm implemented in SIAMCAT.[4] Our classifier achieved an accuracy of 0.868 area under the receiver operating characteristic curve (AUROC) on the discovery cohort, and 0.910 AUROC on the validation cohort (figure 1A). To increase the sample size for model training, we retrained a LASSO classifier for the NKTCL using all the samples from both cohorts, and achieved an accuracy of 0.813 AUROC in cross-validation, which strongly support the role of gut microbiota as diagnostic biomarkers for NKTCL.

To examine the specificity of the NKTCL gut-microbiota-derived signature, we applied the all-sample NKTCL classifier to 29 public gut microbiota cohorts (online supplemental table S1). We observed an overall false positive rate (FPR) of 3.1% in the HCs, but higher FPRs in patients of several cohorts (figure 1B), especially those of the pancreatic cancer (Kartal_DE_2022_PC, Kartal_ES_2022_PC, Nagata_JP_2022_PC), Crohn's Disease (He_2017_CD, Franzosa_2018_CD, Forslund_2015_CD) and liver disease (Qin_2014_LD). These results imply significant overlaps in the biomarkers between these diseases and NKTCL, which was confirmed using LEfSe analysis[5] (figure 1C). Importantly, these biomarkers were consistently enriched/depleted in most cohorts, including the enrichment of oral-derived taxa of *Veillonella* and *Streptococcus* in the patients, and known beneficial species in HCs such as *Faecalibacterium prausnitzii*, *Eubacterium rectale* and *Bifidobacterium adolescentis*[1 6 7] (figure 1C). These findings indicate that our classifier can accurately distinguish NKTCL patients from HCs; nevertheless, due to the shared biomarkers with other diseases, combination of selected clinical indicators with microbial biomarkers would be salutary for a distinctive diagnostic model.

Survival data were available for the NKTCL patients in the discovery cohort. Notably, many identified microbiome biomarkers, especially those shared by multiple diseases, could significantly predict the overall survival (OS) and progression-free survival (PFS) of the patients, including *Streptococcus parasanguinis*, *Romboutsia timonensis* and *Veillonella atypica* (online supplemental figure 1A–D). Finally, we created a *Streptococcus parasanguinis–Romboutsia timonensis* index (SRI) as the relative abundance ratio of the two species, and obtained the best prognostic prediction power than other individual species and combinations. Namely, NKTCL patients with higher SRI scores showed significantly worse OS and PFS than those with lower SRI scores (figure 1D–E). Furthermore, we observed remarkable correlations between high SRI score and multiple adverse prognostic factors of NKTCL, including PINK-E, stage, lymph node

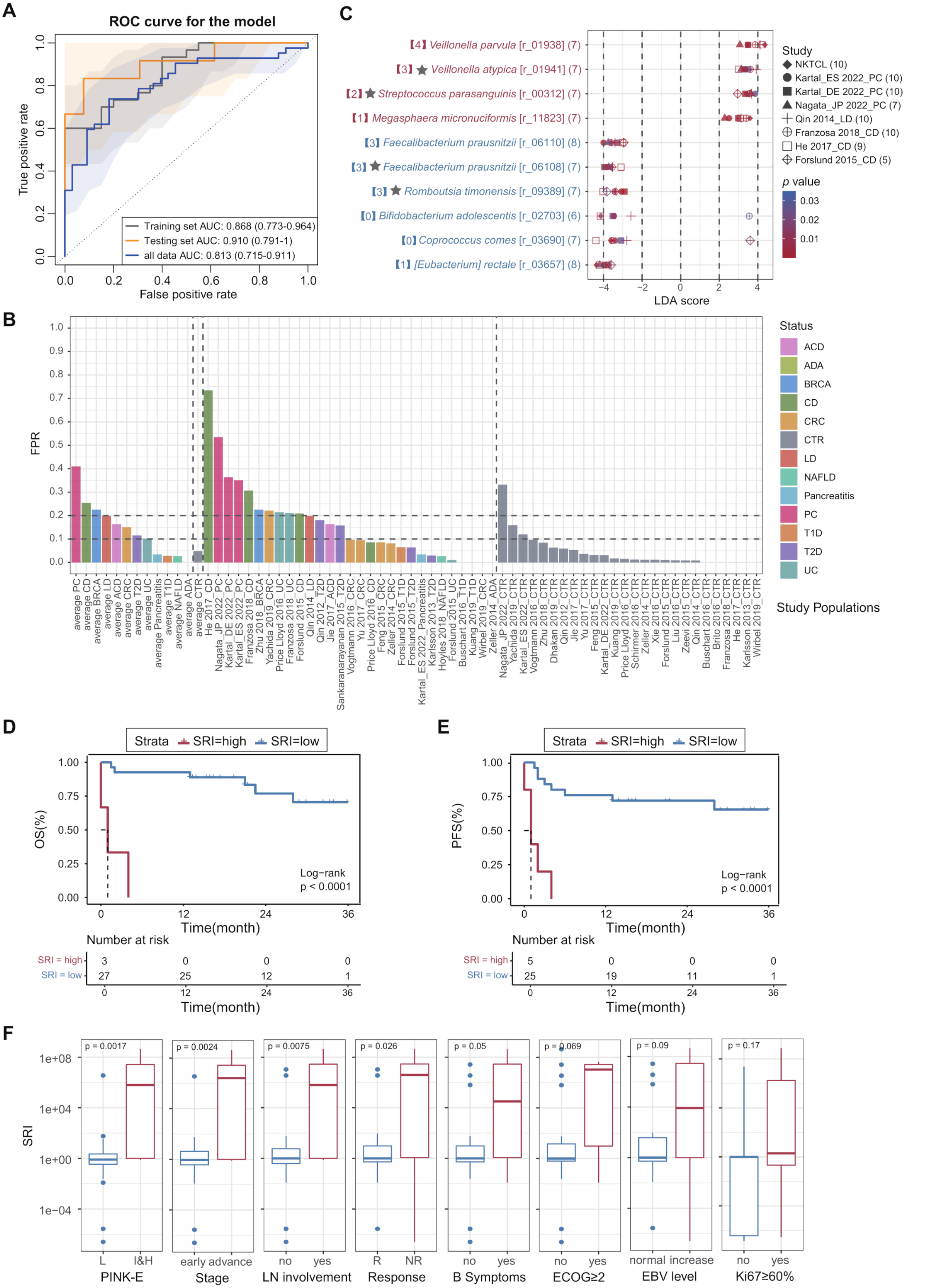


Figure 1 (A) Performance of the area under the receiver operating characteristic curve (AUROC) values of the gut microbiota-based classifier of NKTCL on the discovery cohort (threefold three times repeated cross-validation; grey line, the training set), the validation cohort (yellow line, the testing set), and all samples combined (ten-fold ten times repeated cross-validation; the ‘all data model’, blue line). (B) External validations of the

Figure 1 Continued

disease specificity of the NKTCL faecal microbiota model (the 'all data model'). False positive rates (FPRs) of the unconstrained model (without feature selection) in the 29 external test sets were shown as a bar plot. We defined the false-positive predictions as those wrongly classified as NKTCL by our model. Thus, two FPRs will be calculated for each cohort, one for the healthy controls (ie, the proportion of healthy controls that were wrongly classified as NKTCL), and another for the diseased individuals (ie, the proportion of diseased individuals that were wrongly classified as NKTCL). We then also calculated an overall FPRs for all the healthy controls and each of the diseases. Prediction results from the 'enrichment-constrained' model by selecting NKTCL-enriched biomarkers (enrichment-constrained model) as recommended by Kartal *et al*,[1] were shown in online supplemental figure 1E. (C) Marker microbes shared by the NKTCL cohort and other seven cohorts that had ~20% and higher FPRs in their diseased subjects in (B); markers were identified using the LDA Effect Size (LEfSe) analysis. Red (blue) species name represents its enriched (depleted) in patients. Wilcoxon rank sum test was used to compare the differences in relative abundances between the patients and HCs of the respective cohorts. Inside the square brackets are the numbers of studies in which the species were also among the top features (robustness >50%) of the corresponding disease-stratification classifiers (online supplemental table S2). The 'Star' symbol in front of a species name indicates that the species are significantly associated with patients' survival in our NKTCL cohort; the details can be found in online supplemental figure 1A–D. Inside the parentheses next to the species name is the number of studies in which the corresponding species were identified as a biomarker, that is, with $|LDA|\geq2$. Inside the parentheses after a study name is the total number of species in this figure that were also biomarkers of the study. (D–E) the overall survival (OS) and progression-free survival (PFS) Kaplan-Meier survival curves for NKTCL patients (n=30). Patients were divided into the SRI-high group and SRI-low group according to scores of the *Streptococcus parasanguinis–Romboutsia timonensis* index (SRI), calculated using the quotient of the relative abundances of the two species; the cut-points of SRI 26386550 for OS and 10 776 890 for PFS, and were determined by the 'survminer' R package V.0.4.9[8] (https://github.com/kassambara/survminer). Log-rank test was used to calculate the p values. (F) Correlations between the SRI score and multiple adverse prognostic factors of NKTCL, including prognostic index for natural killer lymphoma-Epstein-Barr virus (PINK-E; L: low risk, I: intermediate risk, H: high risk), disease stage, lymph node (LN) involvement, responses to first-line treatment (R: response, NR: non-response), B symptoms, Eastern Cooperative Oncology Group (ECOG) Performance Status ≥ 2, an increase in plasm Epstein-Barr virus (EBV) DNA level, and Ki67 expression $\geq$60%. Wilcoxon rank sum test was used to compare continuous variables between groups. (More specific descriptions on these results could be found in online supplemental results). ACD, atherosclerotic coronary disease; ADA, American diabetes; BRCA, breast cancer; CD, Crohn's disease; CRC, colorectal cancer; CTR, controls; DE, German; ES, Spanish; JP, Japan; LD, liver disease; NAFLD, non-alcoholic fatty liver disease; PC, pancreatic cancer; T1D, type 1 diabetes; T2D, type 2 diabetes; UC, ulcerative colitis.

involvement and responses to first-line treatment (all $p<0.05$; figure 1F).

Overall, our results lend support for gut microbiota as a potent assistive diagnostic tool for NKTCL. Moreover, the SRI score, based on the shared biomarkers, may have extensive prognostic utility in multiple diseases and deserves further scrutiny (online supplemental discussion).

Zhuangzhuang Shi,[1] Guoru Hu,[2] Min W Li,[2] Lei Zhang,[1,3] Xin Li,[1,3] Ling Li,[1,3] Xinhua Wang,[1,3] Xiaorui Fu,[1,3] Zhenchang Sun,[1,3] Xudong Zhang,[1,3] Li Tian,[1,3] Zhaoming Li,[1,3,4,5] Wei-Hua Chen ,[2,6,7] Mingzhi Zhang[1,3,4]

[1]Department of Oncology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
[2]Department of Bioinformatics and Systems Biology, Huazhong University of Science and Technology College of Life Sciences and Technology, Wuhan, Hubei, China
[3]Lymphoma Diagnosis and Treatment Centre of Henan Province, Zhengzhou, Henan, China
[4]State Key Laboratory of Esophageal Cancer Prevention & Treatment and Henan Key Laboratory for Esophageal Cancer Research, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, Henan, China
[5]Academy of Medical Sciences of Zhengzhou University, Zhengzhou University, Zhengzhou, Henan, China
[6]Institution of Medical Artificial Intelligence, Binzhou Medical University, Yantai, Shandong, China
[7]College of Life Science, Henan Normal University, Xinxiang, Henan, China

Correspondence to Professor Mingzhi Zhang and Professor Zhaoming Li, Department of Oncology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China; mingzhi_zhang1@163.com, fcclizm@zzu.edu.cn and Professor Wei-Hua Chen, Department of Bioinformatics and Systems Biology, Huazhong University of Science and Technology College of Life Sciences and Technology, Wuhan, Hubei, China; weihuachen@hust.edu.cn

Acknowledgements We would like to thank all the clinical doctors from the Lymphoma Diagnosis and Treatment Centre of Henan Province for their kind suggestions, and we also thank all the generous participants of this study for their supports.

Contributors Study concept and design: MZ, W-HC and ZL. Samples collection: ZS, LZ, XL, XW, LL and XF. Data acquisition: ZS, GH, MWL, ZS, ZL, XZ and LT. Analysis and interpretation of data: W-HC, MZ, ZL, ZS, GH and MWL. Technical and material support: LZ, XL, XW, LL, XF, ZS, ZL, XZ and LT. Drafting of the manuscript: GH and ZS. Revising of the manuscript: W-HC, ZL and MZ. All the authors approved the final version of the manuscript.

Funding This work was supported by the National Natural Science Foundation of China (81970184; 82170183; U1904139; 82070209; 82070210).

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was performed in accordance with the Declaration of Helsinki and rules of good clinical practice, and the study was approved by the Ethics Review Committee of the First Affiliated Hospital of Zhengzhou University, Zhenzhou, China. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2022-328256

►

►).

ZS and GH contributed equally.

ZS and GH are joint first authors.

To cite Shi Z, Hu G, Li MW, *et al*. *Gut* 2023;**72**:1999–2002.

Received 11 July 2022
Accepted 30 September 2022
Published Online First 8 November 2022

Gut 2023;**72**:1999–2002. doi:10.1136/gutjnl-2022-328256

ORCID iD
Wei-Hua Chen http://orcid.org/0000-0001-5160-4398

REFERENCES

1 Kartal E, Schmidt TSB, Molina-Montes E, *et al*. A faecal microbiota signature with high specificity for pancreatic cancer. *Gut* 2022;71:1359–72.

2 Shi Z, Zhang M. Emerging roles for the gut microbiome in lymphoid neoplasms. *Clin Med Insights Oncol* 2021;15:117955492110241.

3 Milanese A, Mende DR, Paoli L, *et al*. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 2019;10:1014.

4 Wirbel J, Zych K, Essex M, *et al*. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol* 2021;22:93.

5 Segata N, Izard J, Waldron L, *et al*. Metagenomic biomarker discovery and explanation. *Genome Biol* 2011;12:R60.

6 Montalban-Arques A, Katkeviciute E, Busenhart P, *et al*. Commensal Clostridiales strains mediate effective anti-cancer immune response against solid tumors. *Cell Host Microbe* 2021;29:1573–88.

7 Nagata N, Nishijima S, Kojima Y, *et al*. Metagenomic identification of microbial signatures predicting pancreatic cancer from a multinational study. *Gastroenterology* 2022;163:222–38.

8 Kassambara AKM, Biecek P, Fabian S. *Survminer: Drawing survival curves using 'ggplot2'*, 2021.

Supplementary material of "Gut microbiota as non-invasive diagnostic and prognostic biomarkers for natural killer/T-cell lymphoma"

Supplementary Materials and Methods

Data availability statement

The metagenomic sequencing data reported in this study is available at the China National Center for Bioinformation (CNCB) - National Genomics Data Center (NGDC) under BioProject accession number PRJCA010329. All other data are available in the manuscript including its supplementary files, or from the corresponding authors upon request.

Subjects recruitment

During May 2019 to April 2021, a total of 50 subjects, including 30 treatment-naïve patients with natural killer/T-cell lymphoma (NKTCL) and 20 healthy controls (HCs) were recruited at the First Affiliated Hospital of Zhengzhou University; the patients and controls were matched for age, gender and body mass index (Supplementary table S3). They constituted the discovery cohort for this study. Based on the same criteria, additional 12 treatment-naïve patients with NKTCL and 13 HCs were recruited during October 2021 to February 2022 in the same medical center; they constituted the validation cohort.

All study participants were of Han nationality and lived in the central China, where the typical diet includes wheat flours, rice, vegetables, meat, and beans, etc. All the individuals involved in our study were omnivorous.

Among the participants, NKTCL patients were identified by pathological diagnosis, and the HCs included healthy individuals who visited our hospital for their physical examination. All the enrolled individuals had to meet the following criteria: (1) age 18 years or older; (2) no antibiotics use within two weeks; (3) without any anti-tumor treatments, including chemotherapy, radiotherapy, or immunotherapy; (4) no severe gastrointestinal disorders, including ulcerative colitis, Crohn's disease, or acute diarrhea; (5) no history of severe, progressive or uncontrolled cardiac, hepatic, renal, or mental diseases; (6) no history of drug abuse. Furthermore, all the HCs should meet the following additional criteria, including having the following routine examinations results within the range of healthy individuals, including the physiological and clinical parameters of the blood, urine and stools, liver and kidney function, blood sugar, blood lipid, electrolyte, electrocardiogram and chest X-ray or computed tomography, negative for hepatitis B surface antigen, hepatitis C virus antibody, treponema pallidum antibody and human immunodeficiency virus antibody; in addition, they should have no history of malignant tumor and acute or chronic diseases, including hypertension, diabetes, obesity (BMI ≥30), metabolic syndrome and infectious diseases.

Clinical assessment

Relevant clinical data were captured from consulting physicians, electronic medical records and follow-up systems (Supplementary table S4-5). Survival data for this research was evaluated by overall survival (OS) and progression-free survival (PFS). Here the OS was defined from the date of diagnosis until death from any reason. In addition, PFS was defined as the time of diagnosis until objective disease progression or death from any reason. The stage and risk stratification of patients with NKTCL were based on the Chinese Southwest Oncology Group and Asia Lymphoma Study Group ENKTL (CA) system [1] and the prognostic index for natural killer lymphoma-Epstein-Barr virus (PINK-E) [2], respectively. The 2014 Lugano criteria [3] was used to assess the responses to first-line treatment, and patients were classified as response (R) if they achieved an objective response (complete or partial response), versus non-response (NR) if they had stable disease or progressed on therapy.

Sample collection

Fresh faeces of each study subject were collected in the morning (6:00-9:00 a.m.) using a fecal collection container (Sarstedt, 80.734.311, Germany) and stored without any additives. All collected samples were transferred on ice to a -80°C refrigerator (Haier, DW-86L626, China) within two hours and stored there until the time of DNA extraction.

DNA extraction

Total faecal DNA was extracted using a MagPure Stool DNA KF kit B (Magen, China) according to the manufacturer's instructions. The concentration of genomic DNA in each sample was quantified with a Qubit Fluorometer by using Qubit dsDNA BR Assay kit (Invitrogen, USA) and the quality was checked by running aliquot on 1% agarose gel.

Library construction

1μg DNA was randomly fragmented by Covaris E210, and the fragmented DNA was selected by Magnetic beads (Agencourt, Cat. No. A63882) to an average size of 200-400bp. The selected fragments were through end-repair, 3' adenylated, adapters-ligation, PCR amplifying and the products were purified by the Magnetic beads. The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA were formatted as the final library and qualified by QC. The qualified libraries were sequenced on MGISEQ-2000 platform (BGI-Shenzhen, China) for paired-end sequencing in both directions with a read length of 150 bp (PE-150).

Metagenome data processing

All the raw metagenomic data were trimmed by SOAPnuke v.1.5.2 [4] to remove adapter contamination, low-quality bases, N' base, and duplication. Then the trimmed reads were mapped to the human genome reference sequence (hg19) using Bowtie2 (version 2.3.5.1) [5] with default parameters, and filtered to get clean data.

Taxonomic profiles were obtained using the mOTU profiler v2.5 [6] and filtered to retain species with a relative abundance of ≥10e-5 in ≥2% of the samples.

Modelling and evaluation of a patient stratification classifier to distinguish NKTCL patients from healthy controls

The modeling and evaluation was performed using the SIAMCAT R package v.1.14.0 [7]; R version 4.0 was used throughout the study.

In order to explore the effect of intestinal microbes on the identification of NKTCL patients, we first eliminated taxa from the discovery cohort that had low overall abundance and prevalence (abundance cut-off point: 0.001). Then, features were standardized as centered log-ratio after being normalized by log10 transformation (to prevent endless numbers from the logarithm, a pseudo-count of 1e-05 was applied to all values). Data were randomly split into test and training sets in three times repeated 3-fold cross-validation. The remaining folds were utilized as training data to develop an L1-regularized (LASSO) logistic regression [8] model for each test fold.

The trained metagenomic classifiers for NKTCL were then applied to the validation cohort after applying a data normalization routine, which selected the same set of features and used the same normalization parameters as in the normalization procedure for the discovery cohort.

External validation of NKTCL classifier on 29 public cohorts

To test the specificity of the NKTCL classifier against other disease cohorts, we first combined the samples from both the discovery and validation cohort in order to increase the sample size for model training. We then trained two LASSO classifiers for the NKTCL using all the samples by using the SIAMCAT R package by two methods. The first method is the same as the above, except that data were randomly split into test and training sets in ten times repeated 10-fold cross-validation. In a second approach as Kartal *et al* [9] putting forward, features were filtered by first calculating the single-feature AUROC and then removing features with an AUROC <0.5, thereby selecting features enriched in NKTCL ('enrichment-constrained' model).

To assess the disease specificity of the trained models, we applied the above two models to the 29 gut microbiota studies covering 6,641 samples across twelve diseases used by Kartal *et al* [9] and Nagata *et al* [10] (Supplementary table S1). Within-cohort data normalization was performed according to the instructions by SIAMCAT (the same normalization procedure used for the NKTCL dataset). Then the NKTCL classifiers were applied to these cohorts, which classified the samples as either "healthy" or "NKTCL". The cut-off threshold for the predictions was set to a false-positive rate of 10% among controls in the initial NKTCL study population. Subjects were considered as "false-positives" if they were classified as "NKTCL". Thus, a false-positive rate could be calculated separately for the control and disease groups for each cohort.

Modelling and evaluation of seven public cohorts

To identify the top features whose relative abundances could be used to distinguish the diseased subjects from the controls in seven selected public cohorts including pancreatic cancer (Kartal_DE_2022_PC, Kartal_ES_2022_PC, Nagata_JP_2022_PC), Crohn's Disease (He_2017_CD, Franzosa_2018_CD, Forslund_2015_CD), and liver disease (Qin_2014_LD), we first built a patient-stratification classifier for each of the cohort, by using the same procedures mentioned above, except a 10-fold ten times repeated cross-validation method was used. The top features were defined as those having more than 50% robustness as calculated by the SIAMCAT tool, i.e., the features that were used by ≥50% of the 100 cross-validation models.

Marker identification

We used the linear discriminant analysis effect size (LEfSe) [11] method to identify the marker microbes for each cohort between the control and disease groups. The markers in selected diseases were then compared with those of our NKTCL cohort. The Wilcoxon rank sum test was used to examine whether the relative abundances of the markers were significantly different between the diseased and HC groups within each cohort.

Survival analysis

The survival analysis was performed using the "survminer" R package v.0.4.9 [12], which determined the optimal cut-points to divide the patients into two groups, and evaluated the associations between the marker abundances and patients' survival outcomes. The Kaplan-Meier plot, statistical results and "number at risk" table were also visualized using the "survminer" R package.

Statistical analysis

All statistical analyses, and the analyses involving R packages, were performed in the version 4 of R.

Supplementary Results

Metagenome-based classifiers distinguished NKTCL patients from controls with high accuracy

We constructed three times repeated 3-fold cross-validation LASSO logistic regression models on the discovery cohort, achieving an AUROC of 0.868 (95% CI: 0.773-0.964; figure 1A). The models validated very well on the validation cohort, achieving an AUROC of 0.910 (95% CI: 0.791-1; figure 1A). To increase the sample size for model training, we combined the samples from both the discovery and validation cohorts, built a LASSO logistic regression model, and validated it using ten times repeated 10-fold cross-validation. We achieved an accuracy of 0.813 (95% CI: 0.715-0.911; figure 1A) AUROC for the "all-samples" model, which strongly support the role of gut microbiota as diagnostic biomarkers for NKTCL. The top contributing species to the model are shown in Supplementary figure 1F.

Performance of NKTCL classifier on external cohorts of other diseases

We next tested the specific of our NKTCL classifier on 29 metagenomic datasets (cohorts) of other diseases, including pancreatic cancer, type 1 and type 2 diabetes, colorectal cancer, breast cancer, liver diseases, non-alcoholic fatty liver disease, Crohn's disease and ulcerative colitis (Supplemental table S1). All the datasets contained also healthy controls. In total, these cohorts contained a total of 6,641 samples. Among the cohorts, 28 were used by Kartal *et al* [9]. We added an additional cohort Nagata *et al* [10], which also focused on the pancreatic cancer, similar to the study by Kartal *et al* [9].

By setting a 90% specificity (allowing for 10% false positive predictions) cut-off to our model, we applied it to the samples of the 29 datasets. We defined the false-positive predictions as those wrongly classified as NKTCL by our model. Thus, two false-positive rates (FPRs) will be calculated for each cohort, one for the healthy controls (i.e., the proportion of healthy controls that were wrongly classified as NKTCL), and another for the diseased individuals (i.e., the proportion of diseased individuals that were wrongly classified as NKTCL); please consult the Supplementary table S6 in which we used the CD as an example to demonstrate how the FPRs were calculated. We then also calculated an overall FPRs for all the healthy controls and each of the diseases. As shown in figure 1B, we observed an overall FPR of 3.1% in the healthy controls, suggesting the high specificity of our model against the HCs. However, we observed higher FPRs in patients of several cohorts, especially those of the pancreatic cancer (Kartal_DE_2022_PC, Kartal_ES_2022_PC, Nagata_JP_2022_PC), Crohn's Disease (He_2017_CD, Franzosa_2018_CD, Forslund_2015_CD), and liver disease (Qin_2014_LD).

Performance of "enrichment-constraint" NKTCL classifier on external cohorts

Kartal *et al* also observed high FPRs of their pancreatic cancer (PDAC) classifier on the external cohorts. They resolved this issue by developing a "enrichment-constraint" model, i.e., by selecting features that are significantly enriched in the PDAC patients. Their resulting "enrichment-constraint" model had low FPRs in both the healthy controls and diseases in the 25 external cohorts (see Figure 3 in ref.[9]). To test if their strategy could also work on our dataset, we retrained a classifier using the discovery cohort and the NKTCL-enriched species as the features. We did intra-cohort validation it using three times repeated 3-fold cross-validation and obtained an accuracy 0.812 AUROC (95% CI: 0.689-0.935; Supplementary figure 1G); this model achieved similarly on the validation cohort with a 0.769 AUROC (95% CI: 0.580-0.958; Supplementary figure 1G). We thus also trained a "all sample" model on combined samples of both cohorts using NKTCL-enriched features. This final "enrichment-constrained" model performed quite well in ten times repeated 10-fold cross-validation with an accuracy of 0.864 AUROC (95% CI: 0.776-0.951; Supplementary figure 1G).

We then applied this model to the 29 external cohorts. We observed significantly increased FPRs in both the healthy controls and diseased individuals, suggesting that the “enrichment-constraint” method did not work on our dataset. Our results implied that the NKTCL might be better characterized by both the loss of healthy bacteria and the enrichment of pathogenic bacteria, and both types of bacteria are required to build disease-specific machine learning classifiers.

Marker microbes shared by NKTCL and other diseases

We noticed significant overlap of the microbial biomarkers between the NKTCL cohort and the other disease cohorts, especially those having high FPRs with our classifier (figure 1C). We thus systematically investigated the overlapping of the marker microbes among these cohorts. We observed significant overlap of the markers among these cohorts. Most importantly, these shared biomarkers were consistently enriched/depleted in most cohorts, including the enrichment of oral-derived taxa of *Veillonella* and *Streptococcus* in the patients, and known beneficial species in HCs such as *Faecalibacterium prausnitzii*, *Eubacterium rectale* and *Bifidobacterium adolescentis* (figure 1C). Also, a few biomarkers were also among the top features of the machine-learning classifiers of their respective cohorts, i.e., they were used by ≥50% of the 100 cross-validation models during intra-cohort validation; for example, *Veillonella parvula* is the top feature of multiple models on various cohorts (NKTCL, Kartal_DE 2022_PC, Qin 2014_LD, Franzosa 2018_CD; Supplementary table S2). These findings indicate that due to the shared biomarkers with other diseases, combination of selected clinical indicators with microbial biomarkers would be salutary for a distinctive diagnostic model.

A S. parasanguinis/R. timonensis abundance ratio (SRI score) is of predictive value to NKTCL patient survival, and is based on shared biomarkers

With the availability of the survival data of 30 NKTCL patients in the discovery cohort, we evaluated the prognostic value of the species to patient survival. We identified a total of four species that could significantly predict the overall survival (OS) and/or progression free survival (PFS) of the NKTCL patients, including *Streptococcus parasanguinis*, *Romboutsia timonensis*, *Veillonella atypica*, and *Faecalibacterium prausnitzii* (Supplementary figure 1A-D). We also evaluated the combinations of the four species and created an *S parasanguinis-R.timonensis* index (SRI) as the relative abundance ratios of the two species that had the best performance (figure 1D-E and Supplementary table S7). According to the SRI score, patients were divided into two groups, namely SRI-high and SRI-low at the SRI cut-points of 26,386,550 and 10,776,890 for the OS and PFS, respectively. We observed that the NKTCL patients with higher SRI scores had significantly inferior OS and PFS ($p < 0.001$ for both the OS and PFS; figure 1D-E). In addition, we observed a significant correlation between high SRI score and multiple adverse prognostic factors of NKTCL, including PINK-E, stage, lymph node involvement, and responses to first-line treatment (all $p < 0.05$; figure 1F). Since all the four species are also biomarkers of many diseases (figure 1C), and the SRI index is based on them, we speculate that they can have broad prognostic values in other diseases. In fact, *F. prausnitzii* has been shown to be significantly associated with the patient survival in the PDAC Japan cohort (see Figure 5 in ref.[10]), further supporting our speculation.

Supplementary Discussion

Overall, our results suggest that the gut microbiota could be both of diagnostic and prognostic values for the natural killer/T-cell lymphoma (NKTCL).

Strikingly, there were significant overlaps in the disease biomarkers (i.e., microbial species that show significant differential abundances between the diseased individuals and the non-diseased controls) among the NKTCL and several diseases, including pancreatic cancer (PDAC or PC), liver disease (LD) and Crohn's disease (CD). Although it is not clear for the exact mechanism underlying the cross-disease overlaps, we speculate that the "shared" biomarkers are likely associated with common pathogenic changes of the local gut environments, including inflammation and/or gut epithelial barrier dysfunction, which render the intestinal mucosa more conducive to the same microbial taxa and further account for some overlapped alterations of gut microbiota in different diseases. Furthermore, the NKTCL might be better characterized by both the loss of healthy bacteria and the enrichment of pathogenic bacteria, suggesting both types of bacteria are required to build disease-specific machine learning classifiers, which might be the reason for a higher FPRs in the "enrichment-constrained" diagnostic model than the all-feature model.

In fact, some of the microbial biomarkers also show consistent depletion/enrichment behaviors in multiple other diseases, as shown in our GMrepo v2 database [13] (data repository for human gut microbiota); see the list below for details. Thus, they are less likely caused by diet, life style and living environment, which are often cohort-specific.

- *Veillonella parvula*: https://gmrepo.humangut.info/taxon/29466
- *Veillonella atypica*: https://gmrepo.humangut.info/taxon/39777
- *Streptococcus parasanguinis*: https://gmrepo.humangut.info/taxon/1318
- *Megasphaera micronuciformis*: https://gmrepo.humangut.info/taxon/187326
- *Faecalibacterium prausnitzii*: https://gmrepo.humangut.info/taxon/853
- *Bifidobacterium adolescentis*: https://gmrepo.humangut.info/taxon/1680
- *Coprococcus comes*: https://gmrepo.humangut.info/taxon/410072

Most recently, Priya *et al.* [14] identified a similar set of gut microbes that were shared in patients with colorectal cancer, inflammatory bowel disease and irritable bowel syndrome, including *Peptostreptococcaceae*, *Streptococcus* and *Staphylococcus*. They showed that these biomarkers could impact disease-specific pathophysiological processes through regulation of different host genes. Therefore, studies integrating gut microbiome and host genomics data are urgently needed to unravel the relationships between the "shared" microbial perturbations and the pathogenesis of diverse diseases.

Despite the above results, we feel that our study had several limitations, such as the limited sample sizes, the single-center recruitment strategy, and relatively short observational time for patient survival status. These may affect the generalization ability of our results. Thus, multi-center studies with longitudinal repeated sampling are needed to further validate our findings. In addition, multi-omics survey of the patient samples and mechanism researches with the help of model animals are also needed to provide mechanistical insights underlying the gut microbiome-disease associations.

Supplementary Figure Legends

Supplementary figure 1 (A-D) Associations between selected biomarker species and patient survival outcomes in terms of the overall survival (OS) and progression-free survival (PFS) for NKTCL patients (n = 30). In each analysis, patients were divided into two groups according to the relative abundances; the cut-points were determined automatically by the "survminer" R package v.0.4.9 [12] (https://github.com/kassambara/survminer). The cut-points are: *Streptococcus parasanguinis* [r_00312], 0.001437739 for OS and 0.001077689 for PFS; *Romboutsia timonensis* [r_09389], 7.91e-05 for OS and 1e-10 for PFS; *Veillonella atypica* [r_01941] 0.007384512 for both OS and PFS; and *Faecalibacterium prausnitzii* [r_06108], 0.003101582 both for OS and PFS. (E) External validation results of the disease specificity of the NKTCL "all data" model. False positive rates (FPRs) of the "enrichment-constrained" model by selecting NKTCL-enriched biomarkers (enrichment-constrained model) using methods recommended by Kartal *et al* [9]. ACD, atherosclerotic coronary disease; ADA, American diabetes; BRCA, breast cancer; CD, Crohn's disease; CRC, colorectal cancer; CTR, controls; LD, liver disease; NAFLD, non-alcoholic fatty liver disease; PC, pancreatic cancer; T1D, type 1 diabetes; T2D, type 2 diabetes; UC, ulcerative colitis; JP, Japan; ES, Spanish; DE, German. (F) The heatmap shows the normalized abundance of 11 selected species in the faecal microbiome of the samples. The left panel represents the contribution of each selected feature to the unconstrained model (without feature selection) using all data, and the robustness (the percentage of models in which the feature is included as predictor) of each feature is expressed as a percentage. (G) Performance as the area under the receiver operating characteristic curve (AUROC) values of "enrichment-constrained" diagnostic model on the discovery cohort (three-fold three times repeated cross-validation; grey line, the training set), the validation cohort (yellow line, the testing set), and all samples combined (ten-fold ten times repeated cross-validation; blue line, all data).

Supplementary Tables

Supplementary table S1. External validation cohorts used in this study. The cohort lists, corresponding meta-data, and processed microbial profile data were all obtained from the study by Kartal *et al* [9] and Nagata *et al* [10].

Supplementary table S2. Overlaps between the biomarker species shown in Figure 1C and the top features of the disease-stratification classifiers for selected cohorts. Here the top features of each cohort are those having more than 50% robustness in the corresponding disease-stratification classifier, determined by the SIAMCAT tool (see Materials and Methods for more details).

Supplementary table S3. Participant characteristics at the time of faeces sampling.

Supplementary table S4. Summarized clinical features of patients with natural killer/T-cell lymphoma.

Supplementary table S5. The sequencing depth and meta-information of the samples we collected, including basic information such as age, gender and some clinical features of patients with natural killer/T-cell lymphoma.

Supplementary table S6. CD as an example to demonstrate how the FPRs were calculated. FPR is the number of wrongly classified patients/healthy controls divided by the number of patients/healthy controls.

Supplementary table S7. Evaluations of the prognostic value of the shared species to the survival of NKTCL patients.

References

1 Hong H, Li Y, Lim ST, Liang C, Huang H, Yi P*, et al.* A proposal for a new staging system for extranodal natural killer T-cell lymphoma: a multicenter study from China and Asia Lymphoma Study Group. Leukemia 2020;**34**:2243-8.

2 Kim SJ, Yoon DH, Jaccard A, Chng WJ, Lim ST, Hong H*, et al.* A prognostic index for natural killer cell lymphoma after non-anthracycline-based treatment: a multicentre, retrospective analysis. The Lancet Oncology 2016;**17**:389-400.

3 Cheson BD, Fisher RI, Barrington SF, Cavalli F, Schwartz LH, Zucca E*, et al.* Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2014;**32**:3059-68.

4 Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S*, et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience 2018;**7**:1-6.

5 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods 2012;**9**:357-9.

6 Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M*, et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun 2019;**10**:1014.

7 Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G*, et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biology 2021;**22**:93.

8 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 1996;**58**:267-88.

9 Kartal E, Schmidt TSB, Molina-Montes E, Rodríguez-Perales S, Wirbel J, Maistrenko OM*, et al.* A faecal microbiota signature with high specificity for pancreatic cancer. Gut 2022:gutjnl-2021-324755.

10 Nagata N, Nishijima S, Kojima Y, Hisada Y, Imbe K, Miyoshi-Akiyama T*, et al.* Metagenomic Identification of Microbial Signatures Predicting Pancreatic Cancer From a Multinational Study. Gastroenterology 2022;**163**:222-38.

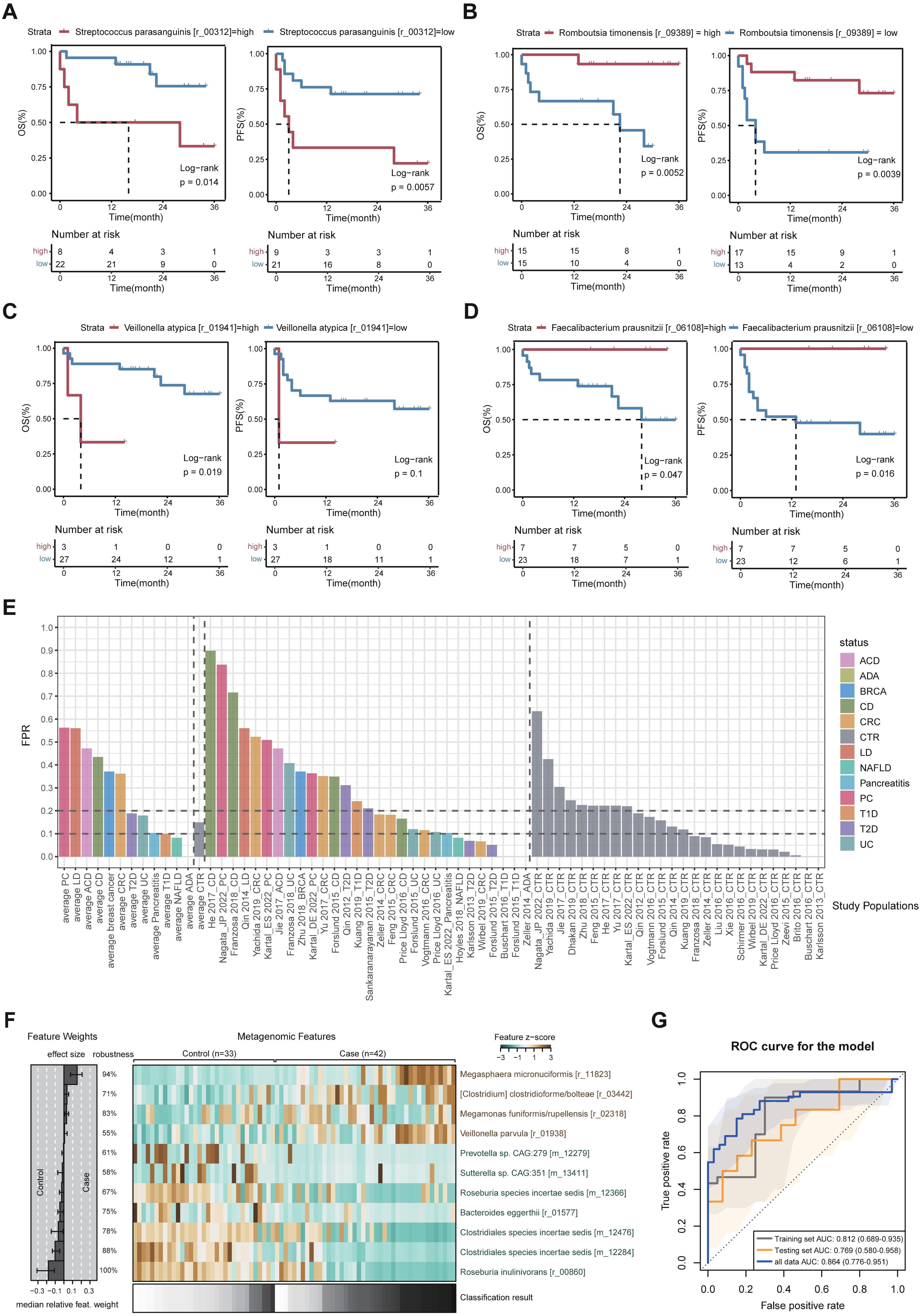
11 Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS*, et al.* Metagenomic biomarker discovery and explanation. Genome Biol 2011;**12**:R60.

12 Kassambara A. KM, Biecek P., Fabian S. Survminer: Drawing survival curves using ‘ggplot2’. 2021-03-09;**version 0.4.9**.

13 Dai D, Zhu J, Sun C, Li M, Liu J, Wu S*, et al.* GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic acids research 2022;**50**:D777-d84.

14 Priya S, Burns MB, Ward T, Mars RAT, Adamowicz B, Lock EF*, et al.* Identification of shared and disease-specific host gene-microbiome associations across human diseases using multi-omic integration. Nat Microbiol 2022;**7**:780-95.

Supplementary figure 1


A
Strata
Streptococcus parasanguinis [r_00312]=high
Streptococcus parasanguinis [r_00312]=low
OS(%)
PFS(%)
Time(month)
Log−rank
p = 0.014
p = 0.0057
Number at risk
high 8 4 3 1
low 22 21 9 0
high 9 3 3 1
low 21 16 8 0
B
Romboutsia timonensis [r_09389] = high
Romboutsia timonensis [r_09389] = low
p = 0.0052
p = 0.0039
high 15 15 8 1
low 15 10 4 0
high 17 15 9 1
low 13 4 2 0
C
Veillonella atypica [r_01941]=high
Veillonella atypica [r_01941]=low
p = 0.019
p = 0.1
high 3 1 0 0
low 27 24 12 1
high 3 1 0 0
low 27 18 11 1
D
Faecalibacterium prausnitzii [r_06108]=high
Faecalibacterium prausnitzii [r_06108]=low
p = 0.047
p = 0.016
high 7 7 5 0
low 23 18 7 1
high 7 7 5 0
low 23 12 6 1
E
FPR
status
ACD
ADA
BRCA
CD
CRC
CTR
LD
NAFLD
Pancreatitis
PC
T1D
T2D
UC
Study Populations
average PC
average LD
average ACD
average CD
average breast cancer
average CRC
average T2D
average UC
average Pancreatitis
average T1D
average NAFLD
average ADA
average CTR
He 2017_CD
Nagata_JP 2022_PC
Franzosa 2018_CD
Qin 2014_LD
Yachida 2019_CRC
Kartal_ES 2022_PC
Jie 2017_ACD
Franzosa 2018_UC
Zhu 2018_BRCA
Kartal_DE 2022_PC
Yu 2017_CRC
Forslund 2015_CD
Qin 2012_T2D
Kuang 2019_T1D
Sankaranarayanan 2015_T2D
Zeller 2014_CRC
Feng 2015_CRC
Price Lloyd 2016_CD
Forslund 2015_UC
Vogtmann 2016_CRC
Price Lloyd 2016_UC
Kartal_ES 2022_Pancreatitis
Hoyles 2018_NAFLD
Karlsson 2013_T2D
Wirbel 2019_CRC
Forslund 2015_T2D
Buschart 2016_T1D
Forslund 2015_T1D
Zeller 2014_ADA
Nagata_JP 2022_CTR
Yachida 2019_CTR
Jie 2017_CTR
Dhakan 2019_CTR
Zhu 2018_CTR
Feng 2015_CTR
He 2017_CTR
Yu 2017_CTR
Kartal_ES 2022_CTR
Qin 2012_CTR
Vogtmann 2016_CTR
Forslund 2015_CTR
Qin 2014_CTR
Kuang 2019_CTR
Franzosa 2018_CTR
Zeller 2014_CTR
Liu 2016_CTR
Xie 2016_CTR
Schirmer 2016_CTR
Wirbel 2019_CTR
Kartal_DE 2022_CTR
Price Lloyd 2016_CTR
Zeevi 2015_CTR
Brito 2016_CTR
Buschart 2016_CTR
Karlsson 2013_CTR
F
Feature Weights
effect size
robustness
Control
Case
median relative feat. weight
Metagenomic Features
Control (n=33)
Case (n=42)
Feature z-score
Megasphaera micronuciformis [r_11823]
[Clostridium] clostridioforme/bolteae [r_03442]
Megamonas funiformis/rupellensis [r_02318]
Veillonella parvula [r_01938]
Prevotella sp. CAG:279 [m_12279]
Sutterella sp. CAG:351 [m_13411]
Roseburia species incertae sedis [m_12366]
Bacteroides eggerthii [r_01577]
Clostridiales species incertae sedis [m_12476]
Clostridiales species incertae sedis [m_12284]
Roseburia inulinivorans [r_00860]
Classification result
G
ROC curve for the model
True positive rate
False positive rate
Training set AUC: 0.812 (0.689-0.935)
Testing set AUC: 0.769 (0.580-0.958)
all data AUC: 0.864 (0.776-0.951)