1

1 Intratumor bacteria is associated with prognosis in clear-cell renal cell carcinoma

- 2 Running title: ITB predicts survival in ccRCC
- 3 Yuqing Li, M.D.^{1#}, Dengwei Zhang, Ph.D.^{2,3#}, Linyi Tan, M.D.¹, Junyao Xu, M.D.⁴, Ting Guo,
- 4 M.D.⁵, Yang Sun, Ph.D.⁶, Rui Zhang, Ph.D.⁷, Yao Cheng, Ph.D.⁷, Haowen Jiang, M.D.^{1*}, Wei
- 5 Zhai, M.D.^{4*}, Yong-xin Li, Ph.D.^{2,3*} and Chenchen Feng, M.D.^{1*}
- ⁶ ¹Department of Urology, Huashan Hospital, Fudan University, Shanghai 200040, China
- ⁷ ²Department of Chemistry and The Swire Institute of Marine Science, The University of Hong
- 8 Kong, Pokfulam Road, Hong Kong SAR, China.
- ³Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Guangzhou,
- 10 China
- ¹¹ ⁴State Key Laboratory of Oncogenes and Related Genes, Department of Urology, Renji
- 12 Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 200127, China
- ⁵Department of Gynecology, Obstetrics and Gynecology Hospital of Fudan University,
- 14 Shanghai 200000, China
- ⁶ Cancer Institute, Xuzhou Medical University, 209 Tongshan Road, Xuzhou, Jiangsu,
- 16 221004, China
- ¹⁷ ⁷Shanghai KR Pharmtech, Inc., Ltd, Shanghai, China
- 18 [#]Equal contributors
- 19 *To whom correspondence may be addressed.
- 20 Emails: urology_hs@163.com (HJ); jacky_zw2002@hotmail.com (WZ); yxpli@hku.hk
- 21 (YxL); fengchenchen@fudan.edu.cn (CF)
- 22 Word count: 5265
- 23
- 24

2

25 ABSTRACT

26 Background

Intratumor bacteria (ITB) plays a role in various cancer types. Its role in clear-cell renal cell
 carcinoma (ccRCC) remains elusive due to small sample size and inadequate decontamination

29 in relevant studies.

30 **Objective**

To establish common and reproducible ITB-associated biomarkers in ccRCC.

32 Design, setting, and participants

This retrospective study comprised seven bulk RNA sequencing datasets from six publicly available cohorts and one in-house Chinese cohort (Renji), one 16S rRNA sequencing dataset from an original Chinese cohort (Huashan), and one publicly available single-cell RNA sequencing dataset. All of these datasets included ccRCC cases.

37 Outcome measurements and statistical analysis

Composition was presented by relative abundance. Overall and progression-free survival were primary outcomes profiled by putative ITB load and risk score, respectively. Potential host interaction was exploratorily analyzed using gene set enrichment analysis and Sparse CCA.

41 **Results and limitations**

Nine cohorts encompassing a total of 1049 ccRCC cases and 130 paired normal tissues were 42 initially analyzed and underwent decontamination. Surprisingly, neither diversity nor 43 composition was differentially distributed between normal and cancer tissue. High putative 44 bacterial load was associated with better overall survival. Notably, a 7-genera dichotomized 45 ITB risk score was indicative of overall survival and a 13-genera dichotomized ITB risk score 46 was predictive of progression-free survival, respectively. Actinomyces, Rothia and 47 Bifidobacterium showed a protective role while Exiguobacterium was a risk factor. A limitation 48 is lack of causation analyses. 49

Conclusions

51	ITB exists in ccRCC. High ITB loads and ITB-risk score predicts better ccRCC survival
52	regardless of sequencing tech, sample processing or racial disparity.
53	
54	KEY WORDS Clear-cell renal cell carcinoma; Intratumor bacteria; Biomarker; Prognosis
55	
56	Patient Summary
57	In this report, we explored the role of intratumor bacteria (ITB) in renal clear-cell carcinoma
58	(ccRCC) in patients with different race and sequencing platforms. Putative ITB load and a 7-
59	genera ITB risk score were associated with overall survival. A 13-genera ITB risk score was
60	predictive of progression-free survival. We conclude that certain ITB features are universally
61	pathogenic to ccRCC.
62	

4

63 INTRODUCTION

Intratumor microorganisms, especially intratumor bacteria (ITB) signature has shown prognostic effect in a variety of cancer types including gastric cancer¹, colorectal cancer², hepatocellular cancer³, etc., establishing microbiome as a novel omics or "second genome" of cancer⁴. However, ITB may vary to vast extents that renders intratumor microbe findings in some cases, hardly reproducible⁵. Amongst all confounders, biomass of subject^{6,7}, race^{8,9}, sequencing tech¹⁰., and contamination¹¹ play the most pivotal roles.

Despite the pitfalls present in ITB studies, true bacterial signatures could be still identified by imperfect sequencing technologies and decontamination processes, which has been demonstrated through experimental validation^{12,13}. Furthermore, when consistent findings emerge from multiple cohorts, the influence of these pitfalls can be minimized, leading to more reliable and robust conclusions.

Clear-cell renal cell carcinoma (ccRCC) is the most common type of malignancy in kidney that is conventionally accepted as sterile organ and ccRCC is expected to harbor a low biomass of ITB. To date, the existing literature on ITB in ccRCC is limited to three full papers¹⁴⁻¹⁶ and one meeting proceeding¹⁷. These studies suffer from small sample sizes, lack of racial diversity, use of a single sequencing technology, and inadequate decontamination process. Consequently, the precise composition and significance of ITB in ccRCC remain elusive. Therefore, a multicohort study focusing on ccRCC is urgently required.

While we highly concur that ITB exists in most, if not all solid tumors including ccRCC, we aim to answer whether common ITB composition exists in ccRCC and whether ITB signature is prognostic, regardless of demographic, racial and sequencing differences. To achieve this, we incorporated various reports on decontamination and multiple ccRCC cohorts encompassing over 1000 cases that vary in region, race, batch, sequencing tech, etc. We aim

- to identify inherent ITB signatures and explore its prognostication in ccRCC in the current 87
- study. 88

6

89 METHODS

90 Study Population

For 16S rRNA sequencing, we retrospectively collected the 217 tumor and 27 normal 91 samples from 217 patients histologically diagnosed clear cell renal carcinoma who underwent 92 partial or radical nephrectomy in Huashan hospital (Shanghai, China) between May 2013 and 93 Oct 2022 under reasonable inclusion criteria (Figure 1). The samples were formalin-fixed and 94 95 paraffin-embedded. We also included 10 negative controls using sliced paraffin from the margin of the block, sampling paraffin only without tissue. Tumor stages were stratified 96 according to the 8th American Joint Committee on Cancer staging system (AJCC)¹⁸ No 97 subjects received preoperative treatments, including immunotherapies or molecular targeted 98 therapies. 99

Six cohorts with the RNA-Seq data available were included in our study. 100 EGAD00001000597¹⁹ as an integrated molecular study of ccRCC and consists of 100 tumor 101 samples. EGAD00001006029 (CheckMate 025; NCT01668784)²⁰ was a prospective clinical 102 trials of the anti-PD-1 antibody nivolumab in advanced clear cell renal cell carcinoma, and 53 103 FFPE tumor tissues were obtained prior to initial therapy for patients enrolled in this study, 104 including 15 patients with the objective response record of immunotherapy. Data from above 105 two datasets (EGAD00001000597 and EGAD00001006029) were requested from the principal 106 strictly via European Genome-phenome Archive (EGA, https://ega-archive.org/) according to 107 the clinical data transfer agreement. GSE102101²¹, GSE126964²², GSE151419²³ were studies 108 concerning on the renal cell carcinoma by organizations located in Singapore, China and 109 Poland, respectively, and the raw sequencing data of tumor and paired normal samples were 110 downloaded from the expression omnibus (GEO) data repository gene 111 (https://www.ncbi.nlm.nih.gov/geo/). Besides, 27 fresh tumor samples with RNA sequencing 112 were supplied by the Renji hospital (Shanghai, China). 113

7

TCGA was a cancer genomics program which molecularly characterized primary cancer and 114 matched normal samples including clear cell renal cell carcinoma (KIRC). However, due to the 115 limit of access to the level 1 or 2 data of TCGA hosted at the Genome Data Commons (GDC) 116 website, the microbiome data processed by by Poore et al. ²⁴ and by Salzberg et al. ²⁵ 117 respectively were directly adopted for downstream bioinformatic analysis in this study. Poore 118 et al. derived the microbiome data from both WGS and RNA-Seq data, and we used the 119 120 normalized and batch effect-corrected data of 532 tumor and 72 paired normal samples. The author performed decontamination in several degrees and got 5 microbial communities 121 122 including data with non-contamination removed (NR), data with likely contaminants removed (LR), data with putative contaminants removed (PR), data with contaminants removed by 123 sequencing "plate-center" combinations (CR), and data with mostly stringent filtering (SR). 124 Salzberg et al. also used the TCGA data but only took WGS data into consideration and shared 125 us with the data including 40 tumor and 35 paired normal samples. Demographics of all cohorts 126 were demonstrated in Table 1. 127

128 16S rRNA gene amplicon sequencing

In preparation for the 16S rRNA gene sequencing, samples were sectioned from the paraffin-embedded 129 130 tissue blocks, which accepted quality testing, purification and nested amplification. To meet the requirement of sufficient DNA for sequencing, the amplified products were detected by DNA electrophoresis, and the 131 eligible samples were kept for further study. 16S rRNA gene sequencing was conducted at the Nonogene 132 Co., Ltd. In brief, genomic DNA was extracted from the tissue samples using the CTAB/SDS method. The 133 16S rDNA V4 region was amplified through PCR employing a primer pair (515F: 5'-134 GTGCCAGCMGCCGCGGTAA-3' and 806R: 5'-GGACTACHVGGGTWTCTAAT-3') with a barcode. 135 Sequencing libraries were prepared using the TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, 136 137 USA) following the manufacturer's instructions. The libraries were subsequently sequenced on the Illumina NovaSeq platform, yielding 250 bp paired-end reads. 138

139 16S rRNA sequencing data processing and analysis

The raw sequencing data of 16S rRNA in FASTQ format underwent processing with QIIME 2 version 140

8

141 2023.2. Quality filtering, denoising, and chimera removal were performed using DADA2, resulting in highquality sequences that were assigned to amplicon sequence variants (ASVs). The feature table was 142 constructed using these ASVs, and taxonomic information was annotated using a Naive Bayes classifier 143 144 trained with the SILVA 138 SSURef NR99 database. ASVs that could not be confidently assigned at the 145 phylum level, as well as non-bacterial ASVs, were excluded from further analysis.

146 To minimize the potential impact of contaminants, we employed a previously established decontamination process. This involved three steps: Firstly, we used the isNotContaminant function in the "decontam" 147 148 algorithm (ref) to identify possible contaminants. This prediction was based on the difference in ASV prevalence between FFPE samples and tissue samples. Secondly, ASVs with a relative abundance greater 149 150 than 0.5% in the FFPE samples were removed. Lastly, ASVs that appeared in less than 5% of the tissue 151 samples were further eliminated to avoid contingency. Only ASVs that met these criteria were retained for 152 downstream analysis.

Bulk RNA sequencing data processing 153

154 Raw RNA sequencing data from tissue samples obtained from six cohorts were acquired online. Sequencing reads were quality-controlled using fastp v0.21.1, with parameters "-1 50 -5 3 -3 3". Filtered reads that were 155 156 shorter than 50 bp were discarded. To quantify human gene expression, the clean reads were aligned to the 157 human reference genome, GRCh38.p13, available in the GENCODE database using HISAT2 v2.2.1. The gene expression values were quantified in transcripts per million (TPM) using StringTie v2.2.1. 158

159 For profiling the intratumor bacteria from bulk RNA-Seq data, clean reads were initially aligned against an indexed database to remove host or contaminant reads. This alignment was performed using bowtie2 v 2.4.5 160 161 with a "--very-sensitive" model. The indexed database included 9 mammalian genomes (hg38, felCat9, canFam6, mm39, rn7, rheMac10, susScr11, galGal6, bosTau9; University of California- Santa Cruz 162 Genome Browser), 2145438 complete bacterial plasmids (PLSDB databse, v.2021 06 23 v2), 13705 163 mitochondrial genomes (NCBI RefSeq database, accessed on Aug 15, 2022), 9443 plastid sequences (NCBI 164 RefSeq database, accessed on Aug 15, 2022), 6093 UniVec sequences (NCBI RefSeq database, accessed on 165 166 Aug 15, 2022), which were considered potential sources of human habitat- or laboratory-associated or extrachromosomal sequence contaminants for taxonomic classification of microbial metagenomic 167 sequences²⁶. Unmapped paired reads were then subjected to KrakenUniq v 1.0.4 for taxonomic assignment 168 169 using a pre-built database. This database includes complete microbial genomes from RefSeq, comprising

9

170 46,711 bacterial genomes, 13,011 viral genomes, and 604 archaeal genomes. Additionally, the database 171 contains 246 eukaryotic pathogens, the UniVec set of standard laboratory vectors, and the GRCh38 human genome. The abundance of bacteria was evaluated at the genus level, which was deemed more accurate than 172 173 the species level.

To ensure the removal of potential false positive assignments, the bacterial genera underwent further 174 175 filtration based on the following criteria: (1) the genus must contain more than 5 reads; (2) number of duplicated kmer must be larger than half of assigned read counts; (3) genome coverage must be larger than 176 177 1e-5. (1) the genus must have a read count greater than 5; (2) the number of duplicated k-mers must exceed half of the assigned read counts; and (3) the genome coverage must be larger than 1e-5. Additionally, efforts 178 179 were made to distinguish the potential host of the identified genera in order to eliminate non-human-180 associated genera that are likely to be contaminants. To accomplish this, information regarding the isolation sources of bacteria deposited in NCBI (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/), 181 IMG/M (https://img.jgi.doe.gov/cgi-bin/m/main.cgi), GOLD (https://gold.jgi.doe.gov/downloads), and BV-182 183 BRC (https://www.bv-brc.org/docs/quick references/ftp.html) was gathered. For identified genera not isolated in these four databases, potential hosts were obtained through a literature search on Google Scholar. 184 185 Based on the available host information, the identified genera were classified into three groups: non-human 186 (genera not isolated from human), human-exclusive (genera exclusively associated with the human host), 187 and mixed (genera derived from either human or other environments). Non-human-associated genera were 188 subsequently excluded from further downstream analysis.

189 **Microbial analysis**

190 For 16S rRNA sequencing data, feature table, taxonomy, and phylogenetic tree after decontamination were combined into a Phyloseq object for downstream processing. To estimate alpha diversity and beta diversity, 191 all samples were rarefied to 2000 sequencing reads. The statistical significance of differences in alpha 192 diversity was assessed by stat compare means function in R package "ggpubr". Difference in microbial 193 compositions was tested using Permutational multivariate analysis of variance (PERMANOVA). 194

195 For the bulk RNA-Seq data, the counts of genera were converted to relative abundance for analysis. Due to 196 the ununiform sequencing depth that would skew the measure of alpha diversity, we did not examine alpha 197 diversity among RNA-Seq data. Rather, we compared the bacterial read counts per million reads, which 198 could provide an indication of bacterial load. The Bray-Curtis dissimilarity among the samples was

10

calculated using the vegdist function in the R package "vegan" and subjected to Principal Coordinates 199 200 Analysis (PCoA). The statistical significance of the findings was evaluated using PERMANOVA analysis with the adonis2 function. 201

202 To evaluate the impact of clinical factors on intratumor microbial communities, a PERMANOVA analysis 203 with 999 permutations was conducted based on Bray-Curtis dissimilarity. To account for multiple 204 comparisons, all P-values were adjusted using the false discovery rate (FDR) method. To explore the relationship between the overall microbial community and overall survival or progression-free survival, 205 206 dimensionality reduction was employed to reduce the complexity of the microbial data. Principal Component Analysis (PCA) was performed using the PCA function in the "FactoMineR" R package. The first five 207 208 principal components (PCs) of the intratumor microbiome PCA were retained to represent the overall intratumor microbiome. Cox proportional hazard regression models were employed to examine the 209 association between each PC and overall survival or progression-free survival. This analysis was conducted 210 using the coxph function in the "survival" package. P-values were adjusted for multiple comparisons using 211 FDR methods. 212

Identifying diagnosis-related microbiome 213

214 For differential abundance testing between tumour and normal tissues in ccRCC, we used relative abundance 215 and counts per million reads (CPM) respectively. We performed Wilcoxon rank-sum tests for each feature in genus level, and corrected the resulting p-values with the BH method. To exclude the bias caused by the 216 217 sample number imbalance, we incorporated only the matched specimen and finally got 24 pairs in Huashan cohort, 10 pairs in GSE102101 (Cohort 3), 11 pairs in GSE126964 (Cohort 4), 13 pairs in GSE151419 218 219 (Cohort 5).

We also used the Random Forest algorithm to further identify the potential features distinguishing the paired 220 samples using randomForest function in the R package " randomForest"²⁷. Ten-fold cross-validation and 221 five repetitions were adopted to help select a specific number of features, whose importance were measured 222 by accuracy and Gini index. 223

224 **Identifying prognosis-related microbiome**

Difference in microbial compositions was first tested between population with long term survival (LTS) and 225 226 short term survival (STS). Due to the inconsistent following months, we used the median survival time in each cohort as the cutoff. Permutational multivariate analysis of variance (PERMANOVA) was used for 227

11

228 testing difference in microbial compositions as above mentioned.

229 Univariate cox was performed to identify the genera whose abundance associated with overall survival and progression free survival. The HR (Hazard Ratio)>1 indicated that the feature was a risk factor for the 230 231 prognosis, while HR<1 indicated that the protect factor. A cluster of genera were preliminarily screened as the input for the least absolute shrinkage and selection operator (LASSO) to exclude the features with 232 potential multi-collinearity. The glmnet function in the R package "glmnet"²⁸ was used and the family was 233 set as "cox" while the other parameters were set default. Finally, we constructed the cox model using coxph 234 235 function in the R package " survival". To fit the model more reasonably, we took the stepwise regression method to help select a formula-based model by Akaike information criterion (AIC). The OS-related risk 236 237 cox model consist of 7 genera, including Abiotrophia, Actinomyces, Bifidobacterium, Dolosigranulum, 238 Faecalibacterium, Kocuria, and Prevotella. The PFS-related risk cox model contained 13 genera, including Acinetobacter, Brachybacterium, Exiguobacterium, Faecalibacterium, Finegoldia, Haemophilus, Kocuria, 239 Lactococcus, Moraxella, Porphyromonas, Prevotella, Rhodococcus, Rothia. The genera with coef>0 in the 240 models were considered risk factors, while those with coef<0 were considered protect factors. Kaplan-Meier 241 survival curves were plotted to report the association between the survival probability and the abundance of 242 243 specific genera. The strategy for grouping included dichotomization of abundance measured by CPM and 244 the presence or not. The significance was examined by log-rank test and two stage hazard rate comparison. Combined with the clinical covariate such as sex, age, tumor stage and grade, the risk score was tested using 245 246 univariate and multivariate cox to determine whether our risk score of microbial features could acted as an independent prognostic factor. 247

We attempted to determine the centrality among the genera involved in the cox model and to find the hub 248 genera. The estimateNetwork function in the R package "bootnet"²⁹ as used and the correlation between the 249 250 features were visualized with the network plot. The influence of each genus was also measured by the indexes including "Strength", "Closeness", "Betweenness" and "ExpectedInfluence". 251

252 Mapping interaction between genera and host gene

253 We previously got the gene expression of 6 cohorts. To filter genes non-related to protein coding, we mapped the gene list to the human genome profile named 'Homo sapiens.GRCh38.109.chr.gtf.gz' downloaded from 254 255 the ENSEMBL website (http://asia.ensembl.org/index.html) and 19142 genes finally remained. The gene expression of TCGA was downloaded from the GDC portal (https://portal.gdc.cancer.gov/) and the data 256

12

257 format was transformed to TPM.

258 To figure out the molecular change, especially the signaling pathway differentiation between the sub-group stratified by the risk score determined by selected microbial features in cox model, we performed the gene 259 set enrichment analysis (GSEA). We used the GSEA function in the R package "clusterProfiler"³⁰, and the 260 KEGG, PID and REACTOME database were all included using R package "msigdbr"³¹. The p-values were 261 262 corrected with the BH method. We took the mantel test to characterize the correlation between interest genera and interest molecular 263 pathways. The mantel test function in the R package "linkET"³² (ref) was used. We dichotomized the genera 264

into two clusters labeled as risk genera and protect genera. To score the immune related function, the single-265 sample gene set enrichment analysis (ssGSEA) method in the R package "GSVA"³³ (ref) was used. The 266 immune cell infiltration was assessed by the quantiseq method using deconvo tme function in the R package 267

"IOBR"³⁴. As there were 15 patients who received the nivolumab immunotherapy and were recorded the 268 objective response rate in the EGAD00001006029 (Cohort 2), we compared the differential genera between 269 270 two groups, that were CB and NCB, using the Chi-Squared test. The prediction ability was adjudged by the area under curve (AUC). 271

272 To macroscopically evaluate the association between tumor microbiome composition and host gene 273 expression, we performed Procrustes analysis. BC dissimilarity was calculated and then the nonmetric multidimensional scaling (NMDS) was used for dimension reduction. The reduced two dimensions or axes 274 275 was input for the rotations and statistical testing in Procrustes analysis. Furthermore, we took the sparse CCA to identify group level correlations between paired host gene expression and microbiome data using the CCA 276 function in the R package "PMA"³⁵. The parameters were set as default. We processed the data before the 277 analysis. The genus whose relative abundance was higher than 0.001 in at least 10% samples were kept, and 278 the data was transformed to the centered log ratio (CLR) format for downstream analysis. We kept the genes 279 whose expression was greater than 0 in half of the samples and then filtered out genes with low variance, 280 using 25% quantile of variance across samples in each disease cohort as cut-off. These filtering resulted in 281 282 a unique microbiome abundance matrix and host gene expression matrix per cohort for downstream analysis, including 12477 gene × 54 taxa in the EGAD00001000597 cohort, 11817 gene × 28 taxa in the 283 EGAD00001006029 cohort, 12633 gene \times 26 taxa in the GSE126964 cohort, 11406 gene \times 26 taxa in the 284 GSE151419 cohort, and 11492 gene \times 60 taxa in the Renji cohort. As the number of tumor samples in 285

13

GSE102101 was small, we didn't performed sparse CCA in this cohort. After the sparse CCA, we got paired 286 287 genus and genes clusters with significant correlation, and they were classified into a component. The genes in each component were implemented with pathway enrichment analysis. The significance was determined 288 by Fisher's exact test and BH method used for adjustment. 289

290 Statistical analysis

291 All data analyses were conducted via RStudio software unless otherwise specified. Visualizations were performed using ggplot2 R package. Two group comparisons were done using Wilcoxon rank-sum test. 292 293 Spearman's correlations were calculated using cor.test function. The heatmap was created using Heatmap in 294 "ComplexHeatmap" R package. In this paper, we used the following notation to indicate the significance levels of P-values: NS (P > 0.05), *0.05 < P < 0.01, **0.01 < P < 0.001, and *** P < 0.001. 295

296 **Transmission Electron Microscopy (EM)**

A total of 20 ccRCC tissue blocks were subject to EM. Fresh tissues were carefully handled immediately 297 after surgical removal. Blocks sliced 1mm³ in size were placed in a culture dish containing an electron 298 microscope fixation solution. Samples were rinsed in 0.1M phosphate buffer (PB, pH 7.4). Samples were 299 then placed at room temperature for 2 hours using 1% osmium tetroxide prepared in 0.1M phosphate buffer 300 301 (PB, pH 7.4). Gradual dehydration was applied, and infiltration was conducted in a mixture of propylene 302 oxide and Epon 812 resin (1:1) at 37°C overnight. Samples were inserted into an embedding mold filled with pure Epon 812 resin. The embedding mold underwent polymerization in a 60°C oven for 48 hours. Ultrathin 303 304 sections (70nm) were cut from resin blocks using an ultramicrotome and placed on 200 mesh Formvarcoated copper grids. Copper grids with sections were stained in a 2% uranyl acetate-saturated alcoholic 305 306 solution for 15 minutes. Following three rinses with ultrapure water, sections were stained with a lead citrate solution for 10 minutes. Copper grid sections were air-dried at room temperature overnight in a copper grid 307 308 box. The grids were observed under a transmission electron microscope (HITACHI, HT7800).

309 **16S rRNA staining**

We performed 16S rRNA staining in 178 samples mounted on a tissue microarray (TMA) chip from the 310 311 Huashan cohort with an established protocol reported by our group previously¹⁹. Briefly, thorough sterilization of hood, blades, and relevant instruments was carried out. Deparaffinized sections were 312 313 dehydrated, and protease K was applied at room temperature. 100 µM of EUB338-cy5 probes (sequence: 5'-GCTGCCTCCCGTAGGAGT-3') diluted in 1 µM working solution were applied and samples were 314

finalized with DAPI (1:500) staining.

RESULTS 318

ccRCC has low-biomass and most ITB are contaminants 319

As most cohorts in this study were sequenced by bulk RNA-seq, we applied a tailored 320 decontamination algorithm (Fig 2A). Analysis of these datasets revealed diverse bacteria 321 present in ccRCC samples (Fig S1). Raw ITB reads took up ~ (1/2.00E-06) of total sequencing 322 reads (Fig S2A) and showed a positive correlation with total reads (Fig S2B). 327 out of 545 323 genera survived after decontamination (Fig S2C). Our passes not only managed to filter out 324 nonhuman reads (Fig S2D) but also showed an increase in proportion of common contaminants 325 326 after decontamination, indicating that some bacteria, previously accepted as contaminants could be indwelling in ccRCC (Fig S2E). Relative abundance of non-human associated 327 bacteria dropped consistently in all cohorts following our decontamination (Fig S2F). Common 328 genera across cohorts after decontamination remained comparable either grouped by dataset or 329 by sample (Fig S3A-D), whereas similar trend for bacterial read drop was noticed in cancer 330 and normal tissue, respectively, further authenticating the remaining reads were true ITB in 331 ccRCC (Fig S3E-F). Compositional atlas demonstrated by relative abundance, as expected, 332 varied drastically across cohorts (Figure S4A-D). Despite so, two phyla, Proteobacteria and 333 334 Firmicutes were present in all bulk-sequenced cohorts (Fig S4) and in scRNA-sequenced samples (Fig S5, Table S1). They were putatively present in diverse cells such as tumor cells 335 (Fig S5). Furthermore, we then applied 16S rRNA-targeted FISH probe and EM imaging to 20 336 ccRCC tissue blocks, validating ITB existence in ccRCC (Fig 2B-C). We also attempted to 337 culture 5 tissue blocks in aerobic and anaerobic conditions, but no bacterial growth was noted, 338 supporting low biomass feature of ccRCC (data not shown). We then cross-referenced top-20 339 abundant genera in all cohorts and found 11 genera were present in \geq 5 cohorts (Fig 2D, Fig 340 S5B, Table S2). Interestingly, three genera including Cutibacterium were also present in 341 TCGA cohort processed by both approaches (Fig S5C). Here, we concluded ccRCC harbored 342

16

a low biomass of ITB and identified presence and composition of ITB which was possibly 343 extracellular in ccRCC. 344

ITB does not differ between adjacent normal and cancer tissue in ccRCC 345

Using decontaminated reads (Table S3), we next probed clinical associations of ITB in three 346 profiles: putative ITB load, ITB signature and individual ITB feature(s). Putative ITB load did 347 not differ between paired normal and cancer tissue in all cohorts (Fig 2E, Fig S6A). Whereas 348 349 cohorts that underwent RNA-seq could not be processed for alpha-diversity, we did not observe a difference in alpha-diversity in the Huashan cohort (Fig 2F, Fig S6B-C). Surprisingly, no 350 351 differences in beta-diversity between normal and cancer tissue were observed in all cohorts (Fig 2G, Fig S6D). The only exception was TCGA P cohort, which was challenged for its 352 overinflated ITB reads (Fig S6E) and the alleged corrected version, TCGA S cohort, again 353 showed no difference (Fig S6F). We thus pursued whether individual ITB feature(s) was 354 differentially distributed and was reproducible. Consistent with barren result of comparison 355 between tumor and normal samples using Wilcoxon Test (Fig S7A-C), although the Random 356 Forest identified 10 candidate differential ITB, this machine learning failed to validate those 357 features with satisfactory predicting efficacy across the cohorts (Fig S7D-E). Again, the 10 358 features showed inconsistent trends in TCGA P cohort and none was significantly different in 359 TCGA S cohort (Fig S8A-B). Here we show astonishingly that, contrary to most studies, 360 differential ITB between adjacent normal and cancer tissue could very well be not present in 361 ccRCC. Our findings highly suggested that most ITB in ccRCC could be inherent intra-tissue 362 bacteria residing in kidney and only individual ITB features altered in abundance in cancer 363 environment, supporting a passenger role of ITB in tumorigenesis stage of the disease. 364

Putative ITB load and risk score predict survival in ccRCC 365

As expected, ccRCC could not be subtyped by ITB signature based on survival (Fig S9A-366

B). Indeed, ITB signature on the whole was not associated with any major clinicopathological 367

17

parameters across cohorts (Fig S10). However, higher putative loads were associated with 368 better overall survival (OS) in three cohorts available with OS profile (Fig 3A). In TCGA S 369 cohort that encompassed a small sample size, higher loads conferred a numerical better survival 370 whereas TCGA P cohort showed no difference, further questioning data processing in 371 TCGA P cohort (Fig S11A). Higher putative loads were solely associated with a better 372 progression-free survival (PFS) in two cohorts, not reproducible in one of our original cohorts 373 374 (cohort 6, Renji) and played a marginally protective role in TCGA S cohort (Fig S11B-C). We then identified the compositional differences between patients with long and short survival, 375 376 and the genera that coexisted and possessed consistent risk in univariate cox across three cohorts was used as input for LASSO and Cox model constructing (Fig S12A-B). The model 377 identified a 7-genera ITB risk score predictive of OS in all three cohorts (Fig 3B, Fig S12C-D, 378 Table S4) but not in either TCGA cohort (Fig S12E, Table S5). Specifically, Actinomyces and 379 Bifidobacterium were protective ITB in ccRCC (Fig S13). Similar methodology was applied 380 to PFS probing and a 13-genera risk score was generated (Fig S14, Table S6). Higher score 381 predicted worsened PFS in all cohorts (Fig 3C) in which *Exiguobacterium* was a risk factor 382 and Rothia was protective (Fig S15). Likewise, the results were not reproducible in either 383 TCGA P or TCGA S cohort (Fig S16, Table S7). Whereas TCGA P was problematic and 384 TCGA S consisted of only WGS samples, we here provided solid evidence that both ITB loads 385 and features played a role in prognosis. This encouraged us to further investigate host 386 interactions and treatment response. 387

ITB is immune-related in ccRCC 388

Of exploratory interest, we investigated interactions between prognosis-related ITB (Fig 389 S12D, Fig S14C) and found Actinomyces and Rothia being consistent hub ITB features across 390 cohorts (Fig S17). When host interactions were incorporated, we found the immune response 391 to be the sole consistently enriched program in ITB risk score-stratified patients across all 392

18

cohorts (Fig 4A, FigS18A-B, Table S8-11). In reminiscence of inter-ITB interactions, ITB 393 genera were associated with antigen presenting cell functions (co-inhibition and co-394 stimulation). (Fig 4B, Fig S18C-D). The risk score ITB showed in general negative correlation 395 with pro-cancer immune infiltrates (Fig 4C). Specifically, absence of protective ITB 396 Actinomyces, Rothia and Bifidobacterium were associated with M2 polarization of 397 macrophages (Fig 4D, Fig S18E-G). Nonetheless, those three features were not associated with 398 response to immune checkpoint inhibition and we identified Anaerococcus and 399 Corynebacterium enriched in ccRCC with complete response (CR) to Nivolumab (Fig 4E-F). 400 401 Lastly, we profiled host interaction using Sparce CCA and three out of five cohorts showed significant host gene-ITB interaction (Fig S19A). Besides immune, we also noted Ribosome 402 signaling was associated with some microbiota across all cohorts (Fig S19B-G). Here, we 403 showed ITB was associated with host immune response in particular protective ITB that were 404 associated with decreased immune escape. 405

19

DISCUSSION 407

Our study encompassed thus far the largest number of ccRCC cases subject to ITB detection. 408 In comparison to previous smaller studies¹⁴⁻¹⁷, several ITB features appeared to be ubiquitously 409 present at high relative abundance including Proteobacteria and Firmicutes at phylum level and 410 Pseudomonas, Acinetobacter and Staphylococcus at genus level. Lack of difference in ITB 411 loads, composition or diversity between normal and cancer tissue was one of our major findings. 412 Though it was previously reported by Wang et al, we initially considered it to be a result of 413 lack of any decontamination in their study¹⁶. Given that ITB features associated with prognosis 414 415 were not amongst the top abundant ones, we speculate that ITB could be sporadic and commensal, not just in ccRCC but also in kidney. 416

Though our 7-genera panel appeared to perform consistently in all cohorts, we are yet to 417 conclude a pathogenic mechanism regarding a single ITB. Like in genetic association studies, 418 prognostic panel composed of multiple genes serves as a biomarker simply because none of 419 the individual gene is statistically powerful enough to generate a reproducible survival 420 difference and any attribution of a single element should be supported by mechanistic analyses 421 by cell or animal modeling. Likewise, our ITB panel solely represents the prognostication of 422 the microbial community. Moreover, our ITB panel was only aggregated at genus rather than 423 species level, further against overinterpretation. 424

The causation between ITB and renal tumorigenesis remains unknown³⁶. Whether those 425 prognostic ITB are still commensal or, playing driving roles alongside tumor progression 426 depends on human microbiota-associated murine models (HMAMMs) and microbe-phenotype 427 triangulation (MPT)³⁶. Unfortunately, there are currently no transgenic murine models for 428 ccRCC³⁷ and culturomics from animal models is therefore inapplicable³⁸. It was surprising that 429 most prognosis-associated ITB features were protective and so were high putative loads, 430 contrary to many oncobiome studies. We did not evaluate absolute ITB loads in our own 431

20

cohorts as loading could not be accurately calculated in transcriptome datasets. However, given 432 that recent study points out that absolute, rather than relative abundance plays more important 433 role in microbiome study³⁹, and that load is prognostic in nasopharyngeal cancer (NPC)⁴⁰, we 434 are now setting up a new line to evaluate association between absolute ITB loads and prognosis. 435 We did not put much effort into imaging ITB. For low-biomass cancer, both LPS and FISH 436 staining could harbor magnified signals from extra-tumor bacterial contamination⁴¹. We 437 consider multiple sequencing platforms together with FISH signaling adequate to prove the 438 existence of ITB. Of note, we did not identify any intracellular bacteria either by EM or scRNA-439 440 seq. This could either be inherent biology of ccRCC or be a result of extreme low biomass of kidney as we successfully identified ITB in all 10 samples of bladder cancer undergoing 441 scRNA-seq in another companion project (data not shown). 442

Recent debate over the landmark cancer microbiome study by Poore et al¹ has drawn much 443 attention in the oncobiome community. In their recent report²⁵, Salzberg's team reasoned two 444 major points that Poore's data should be interpreted with caution, including contamination of 445 human reads into microbial signaling and overinflation of microbial reads by machine learning. 446 We owe great thanks to the Salzberg team for providing us KIRC WGS data processed with 447 their protocol for reproduction and validation of our own findings. Even with the very limited 448 sample size, our model showed a numerical OS prediction. The reason Salzberg's team did not 449 process RNA-seq samples was that they considered poly(A)-based transcriptomes could not 450 capture microbial signals. However, half of our cohorts were poly(A)-based transcriptomic 451 datasets and we were able to retrieve effective reads therein. In fact, most ITB studies using 452 scRNA-seq were also able to capture effective reads given the very few cells compared with 453 bulk sequencing. The "poly(A)" problem in the intratumor microbiome has also been 454 thoroughly discussed^{42,43} and our findings undoubtedly further supported the notion. 455

Last but importantly, we show that certain ITB feature is associated with cancer immunity 456

21

and response to Nivolumab in ccRCC, in reminiscence of recent trial modulation gut 457 microbiome in metastatic ccRCC patients receiving Nivolumab plus ipilimumab therapy⁴⁴. The 458 protective ITB in our findings are closely related to decreased immune escape, e.g., inhibition 459 of antigen presentation and decreased M2 polarization, both showing pro-inflammatory effects. 460 Interestingly, ITB with different clinical associations seldom overlap and we have not 461 identified such an "omnipotent" ITB in ccRCC. Despite so, Corynebacterium is of interest as 462 its abundance ranks top 20 in most cohorts and is associated with Nivolumab response. 463 Bifidobacterium supplement has been shown in trial that augments ICI response in metastatic 464 465 ccRCC patients and our findings that intratumor Bifidobacterium was protective shed light on the thus far elusive mechanism of this gut-tumor asix⁴⁴. We did not analyze ITB in ccRCC 466 patients treated with angiogenesis-targeting therapy though there are a handful of datasets 467 available, as angiogenesis was not amongst the MAMPs we identified (Fig S19B). Given that 468 combination therapy has become the mainstay of metastatic ccRCC treatment, we are now 469 setting up an ITB analysis in such samples. 470

471

CONCLUSION 472

ITB exists in ccRCC. High ITB loads predicted better survival. We also developed a robust 473 ITB score predictive of prognosis regardless of sequencing tech, sample processing or racial 474 disparity. Those parameters and panels serve as novel biomarkers for ccRCC. 475

476

All rights reserved. No reuse allowed without permission.

22

478 **REFERENCES**

 Oosterlinck B, Ceuleers H, Arras W, et al. Mucin-microbiome signatures shape the tumor microenvironment in gastric cancer. *Microbiome*. Apr 21 2023;11(1):86. doi:10.1186/s40168-023-01534-w
 Mouradov D, Greenfield P, Li S, et al. Oncomicrobial Community Profiling Identifies Clinicomolecular

482andPrognosticSubtypesofColorectalCancer.Gastroenterology.2023;165(1):104-120.483doi:10.1053/j.gastro.2023.03.205

484 3. Sun L, Ke X, Guan A, et al. Intratumoural microbiome can predict the prognosis of hepatocellular 485 carcinoma after surgery. *Clinical and Translational Medicine*. 2023;13(7)doi:10.1002/ctm2.1331

486 4. Sepich-Poore GD, Guccione C, Laplane L, Pradeu T, Curtius K, Knight R. Cancer's second genome:
487 Microbial cancer diagnostics and redefining clonal evolution as a multispecies process. *BioEssays*.
488 2022;44(5)doi:10.1002/bies.202100252

489 5. Aykut B, Pushalkar S, Chen R, et al. The fungal mycobiome promotes pancreatic oncogenesis via 490 activation of MBL. *Nature*. 2019;574(7777):264-267. doi:10.1038/s41586-019-1608-2

491 6. Nejman D, Livyatan I, Fuks G, et al. The human tumor microbiome is composed of tumor type–specific
 492 intracellular bacteria. *Science*. 2020;368(6494):973-980. doi:10.1126/science.aay9189

4937.Fu A, Yao B, Dong T, et al. Tumor-resident intracellular microbiota promotes metastatic colonization in494breast cancer. Cell. 2022;185(8):1356-1372.e26. doi:10.1016/j.cell.2022.027

4958.Luo M, Liu Y, Hermida LC, et al. Race is a key determinant of the human intratumor microbiome. Cancer496Cell. 2022;40(9):901-902. doi:10.1016/j.ccell.2022.08.007

497 9. Byrd D, Wolf P. The microbiome as a determinant of racial and ethnic cancer disparities. *Nat Rev Cancer*.
498 Oct 23 2023;doi:10.1038/s41568-023-00638-7

Galeano Niño JL, Wu H, LaCourse KD, et al. INVADEseq to identify cell-adherent or invasive bacteria and
 the associated host transcriptome at single-cell-level resolution. *Nature Protocols*. 2023;doi:10.1038/s41596 023-00888-7

502 11. Dohlman AB, Arguijo Mendoza D, Ding S, et al. The cancer microbiome atlas: a pan-cancer comparative
 503 analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host & Microbe*. 2021;29(2):281 504 298.e5. doi:10.1016/j.chom.2020.12.001

505 12. Colbert LE, El Alam MB, Wang R, et al. Tumor-resident Lactobacillus iners confer chemoradiation
506 resistance through lactate-induced metabolic rewiring. *Cancer Cell*. 2023;41(11):1945-1962.e11.
507 doi:10.1016/j.ccell.2023.09.012

50813.Zhang Z, Gao Q, Ren X, et al. Characterization of intratumor microbiome in cancer immunotherapy. The509Innovation. 2023;4(5)doi:10.1016/j.xinn.2023.100482

51014.Heidler S, Lusuardi L, Madersbacher S, Freibauer C. The Microbiome in Benign Renal Tissue and in Renal511Cell Carcinoma. Urologia Internationalis. 2020;104(3-4):247-252. doi:10.1159/000504029

Liss MA, Chen Y, Rodriguez R, et al. Microbiome within Primary Tumor Tissue from Renal Cell Carcinoma
May Be Associated with PD-L1 Expression of the Venous Tumor Thrombus. *Advances in Urology*. 2020;2020:1-6.
doi:10.1155/2020/9068068

515 16. Wang J, Li X, Wu X, et al. Uncovering the microbiota in renal cell carcinoma tissue using 16S rRNA gene
516 sequencing. *Journal of Cancer Research and Clinical Oncology*. 2020;147(2):481-491. doi:10.1007/s00432-020517 03462-w

518 17. Wheeler C, Yang Y, Spakowicz D, Hoyd R, Li M. 942 The tumor microbiome correlates with response to
519 immune checkpoint inhibitors in renal cell carcinoma. *Journal for ImmunoTherapy of Cancer*. 2021;9(Suppl
520 2):A988-A989. doi:10.1136/jitc-2021-SITC2021.942

Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build
 a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: A Cancer Journal for Clinicians*. 2017;67(2):93-99. doi:10.3322/caac.21388

Sato Y, Yoshizato T, Shiraishi Y, et al. Integrated molecular analysis of clear-cell renal cell carcinoma.
 Nature Genetics. 2013;45(8):860-867. doi:10.1038/ng.2699

Motzer RJ, Tannir NM, McDermott DF, et al. Nivolumab plus Ipilimumab versus Sunitinib in Advanced
 Renal-Cell Carcinoma. *New England Journal of Medicine*. 2018;378(14):1277-1290.
 doi:10.1056/NEJMoa1712126

Yao X, Tan J, Lim KJ, et al. VHL Deficiency Drives Enhancer Activation of Oncogenes in Clear Cell Renal
 Cell Carcinoma. *Cancer Discov*. Nov 2017;7(11):1284-1305. doi:10.1158/2159-8290.CD-17-0375

531 22. Zhao Q, Xue J, Hong B, et al. Transcriptomic characterization and innovative molecular classification of 532 clear cell renal cell carcinoma in the Chinese population. *Cancer Cell Int*. 2020;20:461. doi:10.1186/s12935-020-

23

533	01552-w
534	23. Kajdasz A, Majer W, Kluzek K, et al. Identification of RCC Subtype-Specific microRNAs-Meta-Analysis of
535	High-Throughput RCC Tumor microRNA Expression Data. Cancers (Basel). Feb 1
536	2021;13(3)doi:10.3390/cancers13030548
537	24. Poore GD, Kopylova E, Zhu Q, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic
538	approach. <i>Nature</i> . 2020;579(7800):567-574. doi:10.1038/s41586-020-2095-1
539	25. Gihawi A, Ge Y, Lu J, et al. Major data analysis errors invalidate cancer microbiome findings. <i>mBio</i> .
540	2023;doi:10.1128/mbio.01607-23
541	26. Nakatsu G, Zhou H, Wu WKK, et al. Alterations in Enteric Virome Are Associated With Colorectal Cancer
542	and Survival Outcomes. <i>Gastroenterology</i> . Aug 2018;155(2):529-541 e5. doi:10.1053/j.gastro.2018.04.018
543	27. Liaw, A. and Wiener, M. (2002) Classification and Regression by Randomforest. R News, 2, 18-22.
544	http://CRAN.R-project.org/doc/Rnews/
545	28. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate
546	Descent. Journal of Statistical Software. 2010;33(1)doi:10.18637/jss.v033.i01
547	29. Epskamp S, Borsboom D, Fried EI. Estimating psychological networks and their accuracy: A tutorial
548	paper. Behav Res Methods. Feb 2018;50(1):195-212. doi:10.3758/s13428-017-0862-1
549	30. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.
550	Innovation (Camb). Aug 28 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141
551	31. Dolgalev I (2022). msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format . R
552	package version 7.5.1
553	32. Houyun Huang(2021). linkET: Everything is Linkable. R package version 0.0.7.4.
554	33. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data.
555	BMC Bioinformatics. Jan 16 2013;14:7. doi:10.1186/1471-2105-14-7
556	34. Zeng D, Ye Z, Shen R, Xiong Y (2023). IOBR: Immune Oncology Biological Research . R package version
557	0.99.9.
558	35. Witten D, Tibshirani R (2020). PMA: Penalized Multivariate Analysis . R package version 1.2.1
559	36. Lv B-M, Quan Y, Zhang H-Y. Causal Inference in Microbiome Medicine: Principles and Applications.
560	Trends in Microbiology. 2021;29(8):736-746. doi:10.1016/j.tim.2021.03.015
561	37. van der Mijn JC, Laursen KB, Fu L, et al. Novel genetically engineered mouse models for clear cell renal
562	cell carcinoma. Scientific Reports. 2023;13(1)doi:10.1038/s41598-023-35106-7
563	38. Huang Y, Sheth RU, Zhao S, et al. High-throughput microbial culturomics using automation and machine
564	learning. Nat Biotechnol. Feb 20 2023;doi:10.1038/s41587-023-01674-2
565	39. Maghini DG, Dvorak M, Dahlen A, Roos M, Kuersten S, Bhatt AS. Quantifying bias introduced by sample
566	collection in relative and absolute microbiome measurements. <i>Nature Biotechnology</i> .
567	2023;doi:10.1038/s41587-023-01754-3
568	40. Qiao H, Tan X-R, Li H, et al. Association of Intratumoral Microbiota With Prognosis in Patients With
569	Nasopharyngeal Carcinoma From 2 Hospitals in China. JAMA Oncology.
570	2022;8(9)doi:10.1001/jamaoncol.2022.2810
571	41. de Miranda NFCC, Smit VT, van der Ploeg M, Wesseling J, Neefjes J.
572	2023;doi:10.1101/2023.08.28.555057
573	42. Ghaddar B, Biswas A, Harris C, et al. Tumor microbiome links cellular programs and immunity in
574	pancreatic cancer. Cancer Cell. 2022;40(10):1240-1253.e5. doi:10.1016/j.ccell.2022.09.009
575	43. Hu X, Haas JG, Lathe R. The electronic tree of life (eToL): a net of long probes to characterize the
576	microbiome from RNA-seq data. BMC Microbiology. 2022;22(1)doi:10.1186/s12866-022-02671-2
577	44. Dizman N, Meza L, Bergerot P, et al. Nivolumab plus ipilimumab with or without live bacterial
578	supplementation in metastatic renal cell carcinoma: a randomized phase 1 trial. Nature Medicine.
579	2022;28(4):704-712. doi:10.1038/s41591-022-01694-6
F 0 0	

580

581

24

583 **DECLARATION**

584 Ethics approval and consent to participate

- 585 Informed consent was obtained for all patients and the study was approved by Huashan
- 586 Institutional Review Board (HIRB2011-009; HIRB2023-908) and Renji Hospital, School of
- 587 Medicine, Shanghai Jiao Tong University (KY2023-049-B).

588 Data availability statement

- 589 Read counts of un-decontaminated ITB have been deposited at China National Center for
- 590 Bioinformation (GSA: CRA011414) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa.
- 591 Request for TCGA-KIRC data processed by Salzberg et al should be addressed to Prof. Steven
- 592 Salzberg (steven.salzberg@gmail.com).

593 **Conflict of interest**

594 None.

595 Authors' contributions

596 Conceptualization: CF, WZ, YxL, HJ; Methodology: CF, LT, YL, DZ, YxL; Validation: LT,

597 YL, DZ, HJ; Investigation: CF, LT, YL, DZ, RZ, YS, TG; Original Draft: CF, YL, YC

598 Acknowledgments

This study was sponsored in part by the National Natural Science Foundation of China (Grant No. 81874123 and 82273248). We owe great thanks to the Salzberg team for sharing their data and grant for our use for publication. This study makes use of data generated by the Department of Pathology and Tumor Biology, Kyoto University, and Dr. Seishi Ogawa was highly appreciated. We acknowledge Bristol-Myers Squibb Company (BMS) as the source of the EGAD00001006029 cohort data.

Figure legends 606

Figure 1. Study Design 607

608 Figure 2. The presence of intratumor bacteria in ccRCC

(A) Flowchart illustrating the process of analyzing bulk RNA-seq data to identify intratumor 609 bacteria. The analysis involves using bulk RNA-seq data from normal and tumor tissues. To 610 track potential microbial sources, source annotations from bacterial genomes in databases such 611 as NCBI, IMG/M, GOLD, and BV-BRC, as well as literature search, are retrieved. Bacteria 612 associated with the human host are retained for constructing the intratumor bacteria matrix. (B) 613 614 Representative images of fluorescence in situ hybridization (FISH) staining of 16S rRNA in tumor tissues of ccRCC. (C) Representative images of the presence of bacteria in the tumor 615 tissues captured by transmission electron microscopy based on a total of 20 ccRCC tissue 616 blocks. The red arrow indicates the object. (D) Stacked bar plot showing the proportion of 617 genera present in at least five cohorts among the seven cohorts. Box plot showing the difference 618 of (E) putative load (bacterial counts per million reads) and (F) Shannon index between 24 619 tumor and paired normal samples for comparison in Huashan cohort. The statistically 620 significant difference was given by paired Wilcoxon rank-sum test. (G) Principal coordinate 621 622 analysis (PCoA) for 24 paired tumor and normal samples in Huashan cohort, based on the Bray-Curtis dissimilarity. The P values were tested by Permutational multivariate analysis of 623 variance (PERMANOVA). 624

Figure 3. Putative ITB loads and risk score predict survival in ccRCC 625

Kaplan–Meier curves showing the overall survival probability for Huashan, Cohort 1, and 626 Cohort 2 stratified by (A) putative loads and (B) risk score. (C) Kaplan–Meier curves showing 627 the progression-free survival probability for Huashan, Cohort 2, and Cohort 6 stratified by risk 628 score. P values were calculated using an unadjusted Log-Rank test. 629

Figure 4. ITB is immune-related in ccRCC 630

26

(A) The density curves represent the distribution of the immune-related pathways that were 631 significantly enriched between the two stratified groups using gene set enrichment analysis. 632 The horizontal axis indicated the NES of the GSEA result. The stratification was the same as 633 the previous result, that is the overall survival-related risk group in Cohort 1 and 2, and 634 progression-free survival-related risk group in Cohort 2 and 6 from top to bottom. (B) The 635 result of the Mantel test showing the interaction between genera community and potential 636 immune function and the Spearman method was used. The thickness of the curve indicated the 637 absolute value of the spearman rho, and the significant connection was vellow colored. Each 638 639 block represented the correlation among the immune functions, and a redder color meant a greater rho. (C) Heatmap showing the correlation between specific genus and infiltration scores 640 of immune cells in Cohort 1, Cohort 2, and Cohort 6. (D) Box plot exhibiting the level of M2 641 macrophage polarization in the presence or absence of Actinomyces, Rothia, Bifidobacterium 642 from left to right in Cohort 6. The Wilcoxon Test was used for comparing the relative 643 abundance between tumor and normal. (E) The heatmap at left showed the relative abundance 644 of the differential genera in abundance in the patients with clinical benefit (CB) and non-645 clinical benefit (NCB) using Chi-Squared Test. The heatmap at the right indicated the mRNA 646 expression of genes PDCD1, CD274, and CTLA4. (F) The ability of the abundance of 647 Anaerococcus and Corvnebacterium to predict the clinical benefit was visualized by the 648 receiver operating characteristic curve and measured using the area under the curve (AUC). 649

650

perpetuity. All rights reserved. No reuse allowed without permission.













Genus 1

PCoA1: 17.11 %

medRxiv preprint doi: https://doi.org/10.1101/2023.12.29.23300629; this version posted December 29, 2023. The copyright holder for this perpetuity. Figure 2^{mailor} All rights reserved. No reuse allowed without permission.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.29.23300629; this version posted December 29, 2023. The copyright holder for this perpetuity. Figure S All rights reserved. No reuse allowed without permission.



