

# The Landscape of Microbial Associations in Human Cancer

Abraham Gihawi<sup>1\*</sup>, Henry M Wood<sup>2</sup>, Jeremy Clark<sup>1</sup>, Justin O’Grady<sup>1,3,4</sup>, Rosalind A Eeles<sup>5</sup>, David C Wedge<sup>6,7</sup>, G Maria Jakobsdottir<sup>6,7</sup>, Gkikas Magiorkinis<sup>8</sup>, Andrew G Schache<sup>9</sup>, Liam Masterson<sup>10</sup>, Matt Lechner<sup>11</sup>, Tim R Fenton<sup>12</sup>, Terry M Jones<sup>9</sup>, Adrienne Flanagan<sup>11</sup>, Solange De Noon<sup>11</sup>, Alex Rubinsteyn<sup>13</sup>, Rachel Hurst<sup>1</sup>, Colin S Cooper<sup>1 †</sup>, Daniel S Brewer<sup>1,14 †</sup>

Affiliations:

- 1 Metabolic Health Research Centre, Norwich Medical School,  
University of East Anglia, Norwich, UK, NR4 7UQ
- 2 University of Leeds, UK, LS2 9JT
- 3 Quadram Institute Biosciences, Norwich, UK, NR4 7UQ
- 4 Oxford Nanopore Technologies, UK, OX4 4DQ
- 5 Institute of Cancer Research, London, UK, SW7 3RP  
Manchester Cancer Research Centre University of Manchester,  
UK, M20 4GJ
- 7 NIHR Manchester Biomedical Research Centre, M139WU
- 8 University of Athens, Greece, 106 79
- 9 University of Liverpool, UK, L69 3BX
- 10 Cambridge University Hospitals NHS Trust, UK, CB2 0QQ
- 11 University College London, UK, WC1E 6BT
- 12 University of Southampton, UK, SO17 1BJ  
Lineberger Comprehensive Cancer Center, University of North  
13 Carolina at Chapel Hill, USA, NC 27599
- 14 Earlham Institute, Norwich, UK, NR4 7UZ

\* Corresponding author. Email: A.Gihawi@uea.ac.uk

† These authors jointly contributed to this work

## Abstract:

Oncomicrobes are estimated to cause 15% of cancers worldwide. When cancer whole genome DNA sequencing data (WGS) is collected, microbes present are also sequenced, allowing investigation of potential aetiological and clinical associations. Interrogating the microbial community for 8,908 patients encompassing 22 cancer types from the Genomics England WGS dataset revealed that only colorectal tumours exhibited unmistakably distinct microbial communities that can reliably be used to distinguish anatomical site (PPV=0.95). This pattern was validated in two other datasets. Potential clinical uses uncovered included accurate detection of alphapapillomaviruses (HPV) in oral cancers when compared to current clinical standards, and the detection of rare, highly pathogenic viruses (Human T-Lymphotropic Virus-1). Biomarker investigations demonstrated statistically significant associations ( $P<0.05$ ) between a subset of anaerobic bacteria and survival in certain subtypes of sarcoma. Our results contradict previous claims that each cancer type has a distinct microbiological signature, but highlight the potential value of microbial analysis for certain cancers as WGS of tumour samples becomes common in the clinic.

## Introduction

Well characterised oncomicrobes (1) are attributed with causing 15% of cancers globally (2). These include *Helicobacter pylori* (gastric carcinoma), human papillomavirus (oral, cervical

31 cancer, and others), hepatitis B & C viruses (hepatocellular carcinoma), Epstein-Barr virus  
32 (Hodgkin's lymphoma, Burkitt's lymphoma and nasopharyngeal carcinoma) (2), and HTLV  
33 viruses (Kaposi sarcoma and leukaemias) (3). Specific bacteria such *Fusobacterium*  
34 *nucleatum*, genotoxin-producing *Escherichia coli*, and sets of anaerobic bacteria have been  
35 implicated in colorectal and prostate cancer development, with proposed mechanisms  
36 including DNA damage and immune modulation (4-10).

37

38 Large-scale national sequencing initiatives are leading to the establishment of genomic national  
39 medicine services (11-14). Whole genome sequencing (WGS) of tumour biopsies is likely to  
40 become routine, and its integration into standard clinical care is being considered (15). We  
41 previously used WGS data to survey the landscape of viral associations in human cancer (16)  
42 and have developed SEPATH (17) - a benchmarked approach to identifying microbes in human  
43 tissue WGS data. This approach removes human reads and classifies the remaining reads using  
44 Kraken (17, 18), which has demonstrated applications in clinical diagnostics and surveillance  
45 (19-22). WGS cancer data are considered low-biomass and are challenging to analyse,  
46 particularly distinguishing between biologically relevant and contaminant taxonomic  
47 classifications (23). The latter can arise through various forms of sample contamination as well  
48 as contaminated reference genomes.

49

50 The Cancer Genome Atlas (TCGA) dataset has been investigated for microbial content several  
51 times (23-25). Poore *et al.* (25) investigated microbial classifications in the TCGA dataset  
52 (whole genome and RNA sequencing of blood and cancer samples) and reported that 32 cancer  
53 types exhibited distinct populations of microorganisms with machine learning predictors  
54 giving near-perfect accuracy at distinguishing between cancer types. There were several  
55 surprising findings in this manuscript. Notably, a high total number of sequencing reads were  
56 found in many tumours from sites with no established microbiome, for example glioblastoma..  
57 Classifications of cancer types were also obtained using bacterial sequences in blood, even  
58 though the presence of microbial nucleic acids remains controversial (26-29)

59

60 When re-examining this work, we found two fundamental methodological flaws(30, 31).  
61 First, errors in the processing methods and databases used resulted in millions of DNA  
62 sequence reads being misclassified as microbial across all cancer types. Second, errors in the  
63 methods used to correct batch effects created artificial signatures even when taxa (often  
64 extremophile and nonsensical) were absent in the raw data (30, 31). These observations led  
65 us to conclude that the microbiome classifiers of cancer presented by Poore *et al.* are  
66 incorrect and the article has since been retracted in light of our findings. Nevertheless, the  
67 authors still claim that the cancer microbiome signal is robust over a range of methodological  
68 variation(32), Also a, predominantly theoretical argument has emerged proposing that  
69 sparse/non-existent features becoming associated with disease type may not be evidence of  
70 information leakage (33). Underlying this controversy is that the machine learning models  
71 lack biologically plausible associations and confirmation in independent datasets.

72

73 Here, we investigate the microbial content found within 8,908 patients from 22 different  
74 cancer types within Genomics England's 100,000 Genomes Project sequencing data. This  
75 dataset demonstrates minimal batch effect, circumventing the need for batch correction  
76 approaches. We show that colorectal cancers demonstrate distinctive microbial features and  
77 validate this on two additional datasets (improved classifications of TCGA produced by Ge *et*  
78 *al.* (34) and PCAWG), utilising a total of  $n=21,327$  whole genome sequencing samples to  
79 identify patterns in pancancer microbial structure and potential opportunities for translational

80 benefit. We additionally identify avenues for translational benefit in terms of infectious  
81 disease diagnosis and potential prognostic markers in sarcoma.

82  
83  
84

## 85 **Results**

86 Multiple steps were used to remove potential contamination including human sequence  
87 depletion, confidence thresholding and taxa exclusion. *Homo sapiens* sequences were still  
88 detected in 99.9% of samples despite the use of two methods of depletion (2 to 2,251,317 reads,  
89 median=368, Q1=225, Q3=578). These human counts were excluded as were known common  
90 bacterial contaminants (35) (full list of the genera identified and the taxa removed from  
91 community matrices are provided in table S1 and S2 respectively. All supplementary tables  
92 can be found in data file S1).

93

94 Colorectal and oral cancers are dominated by genera with a high number of sequencing reads  
95 compared to other cancer types. *Bacteroides*, *Parabacteroides*, *Blautia*, *Alistipes* and  
96 *Clostridium* were the most common genera in colorectal cancer, whereas *Prevotella*,  
97 *Fusobacterium*, *Veillonella*, *Actinomyces* and *Gemella* were the most common genera in oral  
98 cancers (figure S1). Clustering of microbial detections revealed limited discernible structure  
99 by tumour site (figure 1). The strongest batch effect involved FFPE status, with weak batch  
100 effects observed for clinical sample geographical location and laboratory sample genomic  
101 medicine centre (figure S2). Biological sex demonstrated a strong split by the number of  
102 unclassified sequencing reads (figure S2G), likely reflecting additional low-complexity regions  
103 within the Y-chromosome. Within FFPE samples, colorectal cancer samples showed a small  
104 grouping, suggesting that there may be some use for identifying microbes in FFPE tissues from  
105 tumours with a higher microbial load. Recognising these variations, we filtered the dataset to  
106 limit these batch effects (for example by removing FFPE and PCR amplified samples) and  
107 curated a list of 495 genera that had potential to be informative of tumour site (table S3).  
108 Clustering the community matrix demonstrated that oral and colorectal microbial communities  
109 contain distinguishing features when compared to other cancer types (Figure 1). 201 genera  
110 were enriched ( $q < 0.05$ , Fisher's exact test with Benjamini-Hochberg Correction) in colorectal  
111 cancer and 114 in oral cancer (Tables S4 and S5, respectively).

112

113

114

## 115 **Elucidating Pan-Cancer Microbial Structure**

116

117 Our finding that only colorectal and oral tumours contain immediately distinctive microbial  
118 communities contrasts previous publications suggesting that the intra-tumoral microbial  
119 community is highly predictive of tumour site (25, 32, 36) including an updated analysis  
120 conducted on partitions of the TCGA data (32). We found that batch effects still exist even  
121 after this partitioning. The metadata features used in batch correction predicted disease type  
122 with high performance (median AUC: 0.975, Q1=0.94, Q3=0.99, 15 models contained PPV  
123 values between 0.99-1, figure S3). Additionally, when partitioning the data by the submitting  
124 centre, a single metadata feature 'tissue source site label' was highly predictive of disease  
125 type (median AUC: 0.92, Q1=0.89, Q3=0.96, figure S4). It is therefore unclear whether high  
126 performance in the updated models(32) is really due to biological signal. We therefore  
127 constructed models in a similar fashion on the Genomics England dataset, with less  
128 observable batch effects (figure 2, S5, S6, S7).

129

130 Generally, our models achieved high AUC values (median: 0.85, Q1=0.79, Q3=0.89), high  
131 specificity (median=0.85, Q2=0.81, Q3=0.96), and reasonable sensitivity (median=0.67,  
132 Q1=0.56, Q3=0.73), but produced comparatively low positive predictive values (PPV; the  
133 probability of disease for a positive test result) (median=0.18, Q1=0.1, Q3=0.34) (figure 2).  
134 The model to predict colorectal cancer samples from all other tumour sites was the only  
135 model to perform significantly better than the negative predictor, with a high PPV of 0.95. It  
136 is noteworthy that the tumour sites with highest positive predictive values are those from  
137 bodily sites with more prominent and widely studied microbial biomass (colorectal, oral,  
138 upper gastrointestinal; PPV=0.95, 0.45, 0.39, respectively). Similar results were observed  
139 with models that were trained on data after applying a read threshold (figure S6) and after  
140 removing the majority of common sequencing contaminants (figure S7). Model feature  
141 importance can be found in table S6.

142  
143 Recently, the microbial composition of tumour samples from the TCGA dataset were profiled  
144 using updated methods revealing a much more sparse community than originally reported  
145 (34). We reanalysed this updated data and found that although there is still a strong batch  
146 effect, the results replicated our finding from the Genomics England cohort: that colorectal  
147 and head and neck tumours (including oral cancer) demonstrate distinctive microbial  
148 communities (figure S8). We identified 85 genera as significantly differentially present in the  
149 TCGA colorectal cohort (Benjamini-Hochberg adjusted Fisher's exact tests,  $q < 0.05$ , table  
150 S7). 69 of these (81%) were also significantly different in the Genomics England cohort  
151 (table S8). Of note, the overlapping genera contained known colorectal constituents as well as  
152 established taxa associated with cancer (for example *Helicobacter* and *Fusobacterim*). The  
153 colorectal cancer result was confirmed in a third cohort, Pan-Cancer Analysis of Whole  
154 Genomes (PCAWG) ( $n=5,041$ ), containing  $n=2,462$  tumour samples. 52 taxa exhibited  
155 differential abundance across all three cohorts (table S9, figure S9-S10). From these  
156 investigations, we conclude that microbial data would only be useful for predicting disease  
157 classification for a restricted set of human cancer types, with only colorectal cancer  
158 exhibiting statistical significance.

159  
160  
161  
162

### 163 **Fungal Genera in Genomics England Dataset**

164  
165 Fungal genera were sparse in the dataset. There was evidence for 113 distinct fungal genera in  
166 the dataset across 6,429 samples. After applying a read threshold of 10, filtering samples to be  
167 PCR-free, non-FFPE primary tumours, only 886 samples remained. 173 samples and 27 fungal  
168 genera had over 100 sequencing reads classified across all samples: *Saccharomyces*,  
169 *Penicillium*, *Enterocytozoon*, *Clavispora*, *Sordaria*, *Fusarium*, *Cyberlindnera*, *Debaryomyces*,  
170 *Nakaseomyces*, *Aspergillus*, *Malassezia*, *Exophiala*, *Botrytis*, *Trichosporon*, *Alternaria*,  
171 *Moesziomyces*, *Meyerozyma*, *Fomitiporia*, *Pseudogymnoascus*, *Rhodotorula*, *Agaricus*,  
172 *Verruconis*, *Purpureocillium*, *Pyrenophora*, *Chaetomium*, *Beauveria*, and *Wickerhamomyces*.  
173 100 of these samples were from colorectal tumours, 17 from lung, 16 from breast, 13 from  
174 sarcoma, 7 ovarian, and 6 renal. The remainder tumour types had fewer than five counts.  
175 Some of these genera may represent environmental or pathobiont species (such as *Aspergillus*  
176 (37) or *Malassezia* (38)) and some may originate from dietary sources (*Saccharomyces* (39)  
177 and *Agaricus* (40)).

178  
179

180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229

## Translational Opportunities for Intratumoural Microbial DNA

We identified several potential clinical uses for identifying the microbial profile from tumour WGS data: Alphapapillomavirus detection that overlaps with somatic tumour features, identification of infectious disease (HTLV-1), and the use of anaerobic bacteria in prognostics.

Head and neck cancer HPV-positive cases represent a distinct disease typically lacking somatic *TP53* mutations and are associated with a favourable prognosis (41). We compared 48 cases of Alphapapillomavirus detection in WGS data against the current gold standard test of mRNA PCR high-risk/tumourigenic subtypes of HPV. The performance using WGS data was excellent, with only one sample not matching the gold standard ( $n=48$ ; sensitivity=100%, specificity=97.3%; Figure 3A). This sample had high HPV burden as detected by WGS and was likely a false negative result for the PCR-based test. As expected, all HPV-positive cases detected as positive (by Kraken or clinical diagnostics) lacked *TP53* mutations (Figure 3). This highlights the use of applying a minimum read threshold for microbial classification using this pipeline, although a threshold of ten may not be optimal for other pipelines.

One participant with invasive breast ductal carcinoma had a total of 172 reads with a Deltaretrovirus classification that were found in tumour and in matching blood samples. We described an ethical framework for reporting highly pathogenic sequences in WGS data and HTLV-1 was identified as a reportable actionable finding (42). All reads in our current analysis uniquely hit HTLV-1 sequences ( $E$ -values  $< 1 \times 10^{-70}$  and percent identities of 100% in all BLAST alignments) with reads across the length of the HTLV-1 reference genome (Figure 3B). These results suggest strong evidence for the computational detection of HTLV-1 in this participant.

In previous work, we identified a set of five bacterial genera associated with aggressive prostate cancer (Anaerobic Bacterial Biomarker Set, ABBS: *Fenollaria*, *Peptoniphilus*, *Anaerococcus*, *Porphyromonas*, *Fusobacterium*) (4). The prostate cohort in Genomics England has limited survival events ( $n=3$ , figure S11). However, within the sarcoma cohort there was a significant association between the presence of at least one ABBS bacteria and survival (log-rank  $P=0.0093$ , figure 3C). This significant association was confirmed in 3/12 sarcoma subtypes and within both genders (figure S12).

## Colorectal Cancer-Specific Microbial DNA in Blood Samples

We investigated our list of recurrent genera specific to colorectal tumours ( $n=52$ ) in blood samples from the PCAWG cohort. Fishers' exact tests for taxa showed that 34/52 (65.4%) were significantly differentially present in blood samples from colorectal patients with cancer compared to blood samples from patients with all other cancer types ( $q<0.05$ , table S10). These genera included *Butyricimonas*, *Parabacteroides*, *Odoribacter*, *Shigella*, *Hungatella*, *Roseburia*, *Porphyromonas*, *Faecalibacterium*, *Blautia*, *Phocaeicola*, *Akkermansia*,

230 *Ruminococcus, Barnesiella, Anaerotignum, Gordonibacter, Bacteroides, Dialister,*  
231 *Clostridioides, Intestinimonas, Flavonifractor, Eubacterium, Parvimonas, Alistipes,*  
232 *Lachnoclostridium, Collinsella, Eggerthella, Anaerostipes, Anaerocolumna, Adlercreutzia,*  
233 *Christensenella, Phascolarctobacterium, Paraprevotella, Megasphaera, and Butyrivibrio .*  
234 These observations indicate that bacterial DNA in the blood may have utility in the diagnosis  
235 of colorectal cancer.

236  
237

## 238 **Discussion**

239

240 In this study we have demonstrated the landscape of microbes that can be identified in  
241 tumour whole genome sequencing data and identified potential translational opportunities  
242 including Alphapapillomavirus assessment, HTLV-1 identification and the potential use of  
243 ABBS genera in sarcoma prognosis.

244

245 We show that oral and colorectal tumours contain distinctive microbial communities. To do  
246 this, we used dimensionality reduction (*t*-SNE), conventional statistics (Fisher's exact tests)  
247 and reconstruction of machine learning models on cleaner datasets than originally published  
248 (tumour types included in different analyses is summarised in table S11) (25). This  
249 observation is replicated in three datasets (Genomics England, TCGA and PCAWG).  
250 Importantly and in contrast to previous analyses (31), the taxa that emerged as differentially  
251 present in colorectal and oral samples generally made biological sense. The results, although  
252 potentially of use in classification, may not have general relevance to cancer development,  
253 with the exception that a small number of known oncomicrobes (*e.g. Helicobacter,*  
254 *Alphapapillomavirus* and *Fusobacterium*) were identified.

255

256 Microbial data in cancer whole-genome sequencing data as completed in our study presents  
257 distinct challenges when compared to conventional microbial analysis. These investigations  
258 are often considered “low biomass” and typically experimental protocols used to generate the  
259 datasets are not specifically designed for microbial investigations (*i.e.* adequate controls,  
260 extraction and sequencing protocols, large proportion of human sequencing reads). There is  
261 also a comparatively high amount of contamination, which can arise from multiple sources  
262 including exogenous (including sequencing reagents, ‘kitome’ and from sites distinct to the  
263 sampling site, *i.e.* patient skin), well-to-well contamination ‘splashome’ (43). These  
264 disproportionately impact low biomass studies, particularly when working with relative  
265 abundance data.

266

267 We have minimised the impact of contamination on our results through various strategies  
268 such as the removal of ubiquitous taxa, the focus on biologically relevant results and the  
269 removal of microbes with low levels of evidence. We provide additional discourse on how  
270 we have mitigated the impact of contamination in our study (supplementary materials and  
271 methods). False positive classifications can arise through contaminated reference genomes.  
272 We would advise the use of curated Kraken databases that have screened genomes for  
273 contamination (such as EuPathDB(44) or GTDB(45)). To mitigate the misclassification of  
274 human reads we include a human reference genome which substantially limits, but does not  
275 entirely remove the misclassification entirely (further discussed in supplementary materials  
276 and methods) (30). As an additional filter, we would expect results from the analyses to make  
277 biological sense, which has not been the case in some studies (31).

278

279 With these improvements only the microbiome present in colorectal cancer can be reliably  
280 used to distinguish between tumour sites. Other cancer types including oral cancer and upper  
281 GI cancers had some distinct microbial features but these did not produce models  
282 significantly better than a negative predictor. While we present robust findings across three  
283 datasets, we for novel observations we advocate the validation of these results using an  
284 orthogonal technology (16S ribosomal sequencing for example). It is important to note that  
285 the TCGA and Genomics England datasets are not always directly comparable. For example,  
286 within TCGA, data is split into colon and rectal, whereas in Genomics England it is grouped  
287 as colorectal. Additionally, in Genomics England, “Upper Gastrointestinal” includes  
288 oesophageal and gastric tumours. Classification performance might have been improved by  
289 separating these subtypes. Cervical cancer is not available in the Genomics England dataset.  
290 Some cancer types were omitted from analyses due to low sample numbers. and despite this,  
291 the key finding that the use of microbiome in the classification of colorectal cancer was  
292 validated in both the PCAWG and TCGA datasets.

293

294 Our results align with the expectation that there is a higher microbial biomass in  
295 oral/colorectal tissue sites compared to other sites that do not hold a known microbial  
296 community (*e.g.* brain), and do not support the existence of a specific ‘cancer microbiome’.  
297 On the application of a minimal read threshold, most taxonomic classifications are removed  
298 from non-oral non-colorectal tumours (figure S13). This is a necessary step to remove many  
299 false positive classifications and we provide an additional description of (this supplementary  
300 materials and methods).

301

302 Some tumour types are well known to have causal associations with the presence of viruses  
303 and bacteria (2). Although they are often causal for a single cancer site, such sequences are  
304 frequently found in multiple locations limiting their use as classifiers for individual cancer  
305 types. This was demonstrated in our previous studies where we examined the landscape of  
306 viruses in human cancer (16). Despite the limited use of microbial composition in  
307 distinguishing cancer types, our results support the clinical utility of using microbial data in a  
308 number of additional specific contexts: in detecting specific viruses such as HPV and HTLV-  
309 1, and in the use of anaerobic bacteria in predicting prognosis.

310

311 Detecting HPV in oral/oropharyngeal carcinoma indicates a distinct biology and is already  
312 used in clinical staging (46). We show here that HPV can be identified at high performance  
313 alongside tumour somatic features with no additional cost. HTLV-1 is a pathogen most  
314 commonly known for causing adult T-cell leukaemia and lymphoma (2). It is a retrovirus that  
315 causes lifelong infections and is predominantly transmitted through breast feeding, sexual  
316 contact, needle sharing and blood transfusions. This highlights how identifying evidence of  
317 infectious disease should be considered as whole genome sequencing increasingly becomes  
318 adopted into clinical practice. Thirdly we identified anaerobic bacteria as a potential  
319 prognostic marker in subtypes of sarcoma. This association is supported by mechanistic  
320 considerations and further research could be done to uncover the exact nature of the  
321 association (4, 47). We also demonstrate that identifying DNA from colorectal-specific  
322 genera in blood samples from colorectal cancer patients could be useful for diagnosing  
323 patients. However, the presence of microbial nucleic acids in blood is controversial (27), and  
324 these results should be validated using an independent cohort. Further research could  
325 establish whether the detected microbial DNA originates from viable microbes or degraded  
326 fragments.

327

328 Overall, our results show that as whole genome sequencing of tumour samples becomes  
329 increasingly used in hospitals, there is potential for the examination of microbial composition  
330 to aid in clinical decisions with no additional financial burden.

331  
332  
333

## 334 **Materials and Methods**

335

### 336 **Study design**

337

338 In this study, the microbial content of  $N=11,735$  human cancer samples from Genomics  
339 England's 100,000 Genomes Project was analysed (48). The aims were to investigate  
340 microbial structure between tumour types and to search for potentially clinically useful  
341 associations. This was carried out with conventional statistics (Fisher's exact tests),  
342 dimensionality reduction approaches and machine learning approaches. Findings were  
343 validated in the PCAWG dataset ( $N=5,041$ , including  $n=2,462$  tumour samples) (16, 49) and  
344 the TCGA dataset ( $N=4,551$ ) (34).

345

346

### 347 **Data**

348 Community matrices, analysis scripts and the reads unmapped to the human genome are  
349 available within the Genomics England research environment for researchers to access. The  
350 community matrix used can be located at the file path:  
351 `/re_gecip/shared_all_GeCIPs/Abe/all_kraken_community.tsv`. Community matrices for the  
352 PCAWG cohort can be found in tables S12-S14 which depict the number of reads, the  
353 number of  $k$ -mers and the coverage of the clade in the database, respectively. The TCGA  
354 reclassifications of Ge *et al.* (34) as used in this manuscript are included as table S15. Users  
355 of these community matrices are strongly advised that they likely contain contamination and  
356 false positive microbial classifications and should be interpreted with caution (31). These  
357 datasets should be used within the context of hypothesis generation and ideally any claims  
358 supported with additional experimental evidence.

359

360

### 361 **Statistical analysis**

362

363 Unless otherwise specified, all statistical analysis was carried out in R (version 4.2.1).  
364 Fisher's exact test was conducted using the `fisher.test` function. Statistical significance was  
365 concluded at  $P<0.05$  (or  $Q<0.05$  for adjusted  $P$ -values). False discovery correction was  
366 carried out using the `p.adjust` function in R using the Benjamini-Hochberg correction  
367 (`method='BH'`). Gradient boosted machine learning models were constructed using scripts  
368 adapted from Poore *et al.* (25). Training-test splits of the data (70% and 30% respectively)  
369 were constructed using the `splitstackshape` R package and stratified by  
370 `'tissue_source_site_label'` for TCGA data partitioned by `'data_submitted_center_label'`.

371

372 For survival analysis, metadata and clinical data was accessed via Rlabkey API within  
373 Genomics England's research environment using release version "main-  
374 programme\_v12\_2021\_05\_06". Date of death was found in either "mortality" or  
375 "death\_details" datasets, which are provided to Genomics England from the Office of  
376 National Statistics and NHS Digital, respectively. For non-deceased participants, date when  
377 they were last seen was inferred from the most recent event from `"hes_ae"` `"hes_apc"`

378 “hes\_cc” “hes\_op” which detail hospital episode statistics from accident and emergency,  
379 admitted patient care, critical care and outpatients respectively. Date of tumour collection  
380 was obtained from the cancer\_analysis dataset. Days to event was calculated as time from  
381 sample collection until date of death or the date the participant was last seen and was divided  
382 by 365 to convert to years. Survival objects were created using the Surv function (survival R  
383 package, version 3.2.3). Survival models were fit with the survfit function (survival R  
384 package) and differences examined using log-rank test. Figures were produced with  
385 ggsvplot function (survminer R package, version 0.4.7). Sarcoma disease subtype was  
386 inferred from disease\_sub\_type of the cancer\_analysis data.

387  
388  
389

## 390 **Taxonomic Classification of Tumour Whole Genome Sequences**

391

392 Samples were collected and processed as per the 100,000 Genomes Project Trial Protocol  
393 (50) and sequenced with the Illumina HiSeq X platform. Sequencing reads were aligned to a  
394 human reference genome (GRCh38) with Illumina iSAAC aligner to produce BAM files.  
395 These BAM files were processed using the SEPATH pipeline (17). In brief, paired-end reads  
396 were extracted if either the forward or the reverse read was unaligned to the human reference  
397 using the PySAM package. These sequencing reads were quality trimmed with Trimmomatic  
398 with parameters: “SLIDINGWINDOW:4:20 MINLEN:35”. The remaining reads were  
399 subject to additional human read depletion using BBDuK (51) using GRCh38, all CDS  
400 sequences in the COSMIC database and additional African human genome variation, with  
401 parameters  $k=30$ ,  $mcf=0.5$  such that at least 50% of the bases in a sequencing read must be  
402 covered by  $k$ -mers present in the reference database for removal. The remaining reads were  
403 subject to taxonomic classification with Kraken (version 1) (18) using a database containing  
404 the human genome (GRCh38) and all bacteria, viral (which includes bacteriophages), fungal  
405 and protozoal genomes at the scaffold level and above (constituent genomes can be found at  
406 <https://zenodo.org/records/15739381>). A confidence threshold of 0.2 was applied to Kraken  
407 reports such that a minimum of 20% of the  $k$ -mers in a sequencing read must be assigned to a  
408 clade for taxonomic classification or the read will remain unclassified.

409  
410

## 411 **Feature Selection and Dimensionality Reduction**

412

413 The sample-taxa Kraken community matrix had a minimum number of 10 reads required for  
414 classification, which appeared to remove a high proportion of classifications with low-level  
415 of evidence (see figure 3A and figure S13). Samples were filtered to represent non-FFPE,  
416 PCR-free, primary tumour samples from cancer types: adult glioma, colorectal, lung,  
417 prostate, bladder, endometrial, malignant melanoma, renal, breast, haematological, oral,  
418 sarcoma, hepatopancreatobiliary, and ovarian. Taxa with total counts across all samples  
419 below 100 were removed from further analysis. Although they may contain biologically  
420 relevant taxa, we removed human classifications and suspected sequencing contaminants  
421 from the community matrices (table S2). This list was informed by investigations into  
422 contamination (35, 52) and ubiquitous presence in the dataset (*Toxoplasma*, *Mycobacterium*,  
423 *Candidatus Pelagibacter*). Although this list may contain biologically relevant taxa, it was  
424 expected that removing these genera would increase biological signal relative to noise  
425 introduced by contamination. *Achromobacter* was also highly prevalent in the dataset but as  
426 ubiquitous as the former three bacteria. It was therefore left in but may resemble  
427 contamination, an opportunistic pathogen or a mixture of both (53). Gradient-boosted

428 machine learning models were constructed to predict the tumour site of a sample compared to  
429 all others for each tumour site individually using scripts provided by Poore *et al.* (25)  
430 (without supervised normalisation). The top 1,500 genera ranked by their feature importance  
431 scores of each model were extracted. The community matrix was further filtered to include  
432 any of the taxa that arose as informative in this feature selection. 495 microbial genera  
433 remained after this filtering (table S3). Of the remaining samples and remaining taxa, a  
434 distance matrix was constructed using the distanceMatrix function (ClassDiscovery R  
435 package). The distance matrix was subject to *t*-SNE (Rtsne R package) with parameters:  
436 dims=2, perplexity=80, max\_iter=2000, check\_duplicates=TRUE.

437  
438

### 439 **Mutation Calling and Analysis**

440

441 All Genomics England somatic genomic samples have a matched germline, sequenced at  
442 100x and 30x respectively. Samples were sequenced with 150bp paired-end reads in a single  
443 lane of Illumina HiSeq X and processed by the illumina North Star Version 4 Whole Genome  
444 Sequencing Workflow (NSV4, version 2.6.53.23). The workflow uses iSAAC Aligner  
445 (version 03.16.02.19)(54) against the *Homo Sapiens* NCBI GRCh38 assembly with decoys  
446 and the small variant caller Strelka2 (version 2.4.7) (55), which performs a probabilistic  
447 subtraction of tumour-normal for the somatic calls. SNVs and indels were then annotated  
448 using CellBase, an in-house tool with more than 99% agreement with the Ensembl VEP  
449 Consequence type. Non-synonymous variants of moderate or high impact, according to the  
450 Ensembl variant consequence list, were investigated in oral/oropharyngeal cohort. These  
451 were identified by using functions provided by Genomics England (01.functions.R) available  
452 within Genomics England's research environment. These functions compile the variants for a  
453 given gene across the cohort. Small gene variants of moderate or high impact were  
454 determined by the following consequence types: transcript ablation, splice acceptor variant,  
455 splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript  
456 amplification, inframe insertion, inframe deletion, inframe variant, missense variant, splice  
457 region variant. Samples with no identified small variants were considered wild-type.

458  
459

### 460 **Clinical HPV Diagnostics**

461

462 The diagnostic pathway for oropharyngeal cases involved routine testing for p16 by  
463 immunohistochemistry. Samples were labelled HPV-positive if p16+ only (as this has  
464 been accepted as a robust proxy measure for HPV status).

465  
466

467

### 468 **HTLV-1 Investigation**

469

470 Participants demonstrating fewer than 20 genus level reads for each of the infectious agents  
471 described in Magiorkinis *et al.* 2019 (42) (HIV, HBV, HCV, HTLV-1) were considered false  
472 positive classifications. Only one participant in the cohort was identified as positive for  
473 HTLV-1. In total, 172 sequencing reads from the tumour and germline sample with any  
474 Deltaretrovirus classification as reported by Kraken were extracted and subject to a BLASTn  
475 (56) via the online suite with standard databases (nr/nt nucleotide collection) optimised for  
476 highly similar sequences (megablast). The query reads from both samples were aligned to

477 HTLV-1 reference genome (NC\_001436.1) using BWA-MEM (57) with standard parameters  
478 which was subsequently visualised with IGV version 2.9.4 (58).

479  
480  
481

## 482 **Supplementary Materials:**

483  
484 Figures S1-13  
485 Tables S1-15 (in data file S1)  
486 MDAR checklist  
487 Supplementary Materials and Methods

488

## 489 **Acknowledgements**

490

491 The authors would like to thank Mariana Buongiorno Pereira for developing template  
492 scripts for survival analysis within the Genomics England research environment. This  
493 research was made possible through access to data in the National Genomic Research  
494 Library, which is managed by Genomics England Limited (a wholly owned company of the  
495 Department of Health and Social Care). The National Genomic Research Library holds data  
496 provided by patients and collected by the NHS as part of their care and data collected as part  
497 of their participation in research. The National Genomic Research Library is funded by the  
498 National Institute for Health Research and NHS England. The Wellcome Trust, Cancer  
499 Research UK and the Medical Research Council have also funded research infrastructure.  
500 Thank you to the participants and families that have made this research possible.

501  
502  
503

504 **Funding:** This work was funded by the Big C Cancer Charity (ref 16-09R, recipient: DSB)  
505 and Prostate Cancer UK (research grant ref: MA-ETNA19-003 recipient: DSB, RIA15-ST2-  
506 029 recipients: DSB & CSC and TLD-CAF22-011 recipient: AG). We are grateful for and  
507 acknowledge support from The Masonic Charitable Foundation Successor to The Grand  
508 Charity, Movember, The Prostate Cancer Research, The King Family and the Stephen  
509 Hargrave Trust (recipient: CSC).

510  
511

## 512 **Author Contributions**

513

514 AG is responsible for Methodology, Formal Analysis, Investigation, Curated Data, Writing  
515 the Original Draft, Review & Editing, Data Visualisation and Project Management.

516 SDN curated data on sarcoma subtypes.

517 AG, CSC and DSB Conceptualized the study design and acquired funding for the project and  
518 are responsible for the interpretation of study data.

519 RH, DSB and CC are responsible for supervision and project management.

520 AGS, LM, ML, TRF TMJ and HMW advised on the analysis an interpretation of study data  
521 (Alphapapillomavirus).

522 AR advised on formal analysis (machine learning classifiers on metadata)

523 AG, HMW, JC, JOG, RAE, DCW, GMJ, GM, AGS, LM, ML, TRF, TMJ, AF, SDN, AR,  
524 RH, CSC, DSB reviewed drafts of the manuscript and helped critique for important  
525 intellectual content.

526

527 **Competing Interests**

528

529 Colin S. Cooper, Daniel S. Brewer, Rachel Hurst, Abraham Gihawi, and Justin O'Grady are  
530 coinventors on a patent application (UK Patent Application No. 2200682.9) from the  
531 University of East Anglia/UEA Enterprises Limited regarding the application of ABBS  
532 genera in prostate cancer. Justin O'Grady is an employee of Oxford Nanopore Technologies  
533 and holds stock and stock options in the company and has previously received honoraria from  
534 Oxford Nanopore.

535

536 **Data and materials availability**

537 All data associated with this study are in the paper or supplementary materials. The Kraken  
538 database consisting of GRCh38 and all bacteria, viral (which includes bacteriophages), fungal  
539 and protozoal genomes at the scaffold level and above (constituent genomes can be found at  
540 <https://zenodo.org/records/15739381>). Community matrices, analysis scripts and DNA reads  
541 unmapped to the human genome are available within the Genomics England research  
542 environment for researchers to access. The community matrix used can be located at the file  
543 path: /re\_gecip/shared\_all\_GeCIPs/Abe/all\_kraken\_community.tsv.

544

545

546

547

548

549

550 **References**

551

- 552 1. S. Garrett, Cancer and the microbiota. *Cancer Immunol Immunother* **348**, 80-86  
553 (2015).
- 554 2. M. Plummer, C. de Martel, J. Vignat, J. Ferlay, F. Bray, S. Franceschi, Global burden of  
555 cancers attributable to infections in 2012: a synthetic analysis. *The Lancet Global*  
556 *Health* **4**, e609-e616 (2016).
- 557 3. P. H. Goncalves, T. S. Uldrick, R. Yarchoan, HIV-associated Kaposi sarcoma and  
558 related diseases. *AIDS* **31**, 1903-1916 (2017).
- 559 4. R. Hurst, E. Meader, A. Gihawi, G. Rallapalli, J. Clark, G. L. Kay, M. Webb, K. Manley,  
560 H. Curley, H. Walker, R. Kumar, K. Schmidt, L. Crossman, R. A. Eeles, D. C. Wedge, A.  
561 G. Lynch, C. E. Massie, C.-I. P. Group, M. Yazbek-Hanna, M. Rochester, R. D. Mills, R.  
562 F. Mithen, M. H. Traka, R. Y. Ball, J. O'Grady, D. S. Brewer, J. Wain, C. S. Cooper,  
563 Microbiomes of Urine and the Prostate Are Linked to Human Prostate Cancer Risk  
564 Groups. *Eur Urol Oncol*, (2022).
- 565 5. P. Georgeson, R. S. Steinfeldt, T. A. Harrison, B. J. Pope, S. H. Zaidi, C. Qu, Y. Lin, J. E.  
566 Joo, K. Mahmood, M. Clendenning, R. Walker, E. K. Aglago, S. I. Berndt, H. Brenner,  
567 P. T. Campbell, Y. Cao, A. T. Chan, J. Chang-Claude, N. Dimou, K. F. Doheny, D. A.  
568 Drew, J. C. Figueiredo, A. J. French, S. Gallinger, M. Giannakis, G. G. Giles, E. L.  
569 Goode, S. B. Gruber, A. Gsur, M. J. Gunter, S. Harlid, M. Hoffmeister, L. Hsu, W. Y.  
570 Huang, J. R. Huyghe, J. E. Manson, V. Moreno, N. Murphy, R. Nassir, C. C. Newton, J.  
571 A. Nowak, M. Obon-Santacana, S. Ogino, R. K. Pai, N. Papadimitrou, J. D. Potter, R. E.  
572 Schoen, M. Song, W. Sun, A. E. Toland, Q. M. Trinh, K. Tsilidis, T. Ugai, C. Y. Um, F. A.  
573 Macrae, C. Rosty, T. J. Hudson, I. M. Winship, A. I. Phipps, M. A. Jenkins, U. Peters, D.  
574 D. Buchanan, Genotoxic colibactin mutational signature in colorectal cancer is

- 575 associated with clinicopathological features, specific genomic alterations and better  
576 survival. *medRxiv*, (2023).
- 577 6. C. Pleguezuelos-Manzano, J. Puschhof, A. Rosendahl Huber, A. van Hoeck, H. M.  
578 Wood, J. Nomburg, C. Gurjao, F. Manders, G. Dalmasso, P. B. Stege, F. L. Paganelli,  
579 M. H. Geurts, J. Beumer, T. Mizutani, Y. Miao, R. van der Linden, S. van der Elst, C.  
580 Genomics England Research, K. C. Garcia, J. Top, R. J. L. Willems, M. Giannakis, R.  
581 Bonnet, P. Quirke, M. Meyerson, E. Cuppen, R. van Boxtel, H. Clevers, Mutational  
582 signature in colorectal cancer caused by genotoxic pks(+) *E. coli*. *Nature* **580**, 269-  
583 273 (2020).
- 584 7. K. Hoppe-Seyler, F. Bossler, J. A. Braun, A. L. Herrmann, F. Hoppe-Seyler, The HPV  
585 E6/E7 Oncogenes: Key Factors for Viral Carcinogenesis and Therapeutic Targets.  
586 *Trends Microbiol* **26**, 158-168 (2018).
- 587 8. C. Gur, Y. Ibrahim, B. Isaacson, R. Yamin, J. Abed, M. Gamliel, J. Enk, Y. Bar-On, N.  
588 Stanietsky-Kaynan, S. Copenhagen-Glazer, N. Shussman, G. Almogy, A. Cuapio, E.  
589 Hofer, D. Mevorach, A. Tabib, R. Ortenberg, G. Markel, K. Miklic, S. Jonjic, C. A.  
590 Brennan, W. S. Garrett, G. Bachrach, O. Mandelboim, Binding of the Fap2 protein of  
591 *Fusobacterium nucleatum* to human inhibitory receptor TIGIT protects tumors from  
592 immune cell attack. *Immunity* **42**, 344-355 (2015).
- 593 9. J. Abed, J. E. Emgard, G. Zamir, M. Faroja, G. Almogy, A. Grenov, A. Sol, R. Naor, E.  
594 Pikarsky, K. A. Atlan, A. Mellul, S. Chaushu, A. L. Manson, A. M. Earl, N. Ou, C. A.  
595 Brennan, W. S. Garrett, G. Bachrach, Fap2 Mediates *Fusobacterium nucleatum*  
596 Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc.  
597 *Cell Host Microbe* **20**, 215-225 (2016).
- 598 10. A. D. Kostic, E. Chun, L. Robertson, J. N. Glickman, C. A. Gallini, M. Michaud, T. E.  
599 Clancy, D. C. Chung, P. Lochhead, G. L. Hold, E. M. El-Omar, D. Brenner, C. S. Fuchs,  
600 M. Meyerson, W. S. Garrett, *Fusobacterium nucleatum* potentiates intestinal  
601 tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host*  
602 *Microbe* **14**, 207-215 (2013).
- 603 11. C. Turnbull, Introducing whole-genome sequencing into routine cancer care: the  
604 Genomics England 100 000 Genomes Project. *Ann Oncol* **29**, 784-787 (2018).
- 605 12. F. Lethimonnier, Y. Levy, Genomic medicine France 2025. *Ann Oncol* **29**, 783-784  
606 (2018).
- 607 13. N. Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R.  
608 Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, The Cancer  
609 Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).
- 610 14. H. Zayed, The Arab genome: Health and wealth. *Gene* **592**, 239-243 (2016).
- 611 15. B. Rahman, A. Lamb, A. Protheroe, K. Shah, J. Solomons, J. Williams, E. Ormondroyd,  
612 Genomic sequencing in oncology: Considerations for integration in routine cancer  
613 care. *Eur J Cancer Care (Engl)* **31**, e13584 (2022).
- 614 16. M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, N. Desai, H.  
615 Sultmann, H. Moch, P. Pathogens, C. S. Cooper, R. Eils, V. Ferretti, P. Lichter, P.  
616 Consortium, The landscape of viral associations in human cancers. *Nat Genet* **52**,  
617 320-330 (2020).
- 618 17. A. Gihawi, G. Rallapalli, R. Hurst, C. S. Cooper, R. M. Leggett, D. S. Brewer, SEPATH:  
619 benchmarking the search for pathogens in human tissue whole genome sequence  
620 data leads to template pipelines. *Genome Biol* **20**, 208 (2019).

- 621 18. D. Wood, S. Salzberg, Kraken - Ultrafast Metagenomic Sequence Classification Using  
622 Exact Alignments. *Genome Biol* **15**, (2014).
- 623 19. C. Smith, T. A. Halse, J. Shea, H. Modestil, R. C. Fowler, K. A. Musser, V. Escuyer, P.  
624 Lapiere, Assessing Nanopore Sequencing for Clinical Diagnostics: a Comparison of  
625 Next-Generation Sequencing (NGS) Methods for Mycobacterium tuberculosis. *J Clin*  
626 *Microbiol* **59**, (2020).
- 627 20. C. Grumaz, A. Hoffmann, Y. Vainshtein, M. Kopp, S. Grumaz, P. Stevens, S. O. Decker,  
628 M. A. Weigand, S. Hofer, T. Brenner, K. Sohn, Rapid Next-Generation Sequencing-  
629 Based Diagnostics of Bacteremia in Septic Patients. *J Mol Diagn* **22**, 405-418 (2020).
- 630 21. R. Yee, F. P. Breitwieser, S. Hao, B. N. A. Opene, R. E. Workman, P. D. Tamma, J. Dien-  
631 Bard, W. Timp, P. J. Simner, Metagenomic next-generation sequencing of rectal  
632 swabs for the surveillance of antimicrobial-resistant organisms on the Illumina Miseq  
633 and Oxford MinION platforms. *Eur J Clin Microbiol Infect Dis* **40**, 95-102 (2021).
- 634 22. S. L. Salzberg, F. P. Breitwieser, A. Kumar, H. Hao, P. Burger, F. J. Rodriguez, M. Lim,  
635 A. Quinones-Hinojosa, G. L. Gallia, J. A. Tornheim, M. T. Melia, C. L. Sears, C. A.  
636 Pardo, Next-generation sequencing in neuropathologic diagnosis of infections of the  
637 nervous system. *Neurol Neuroimmunol Neuroinflamm* **3**, e251 (2016).
- 638 23. A. B. Dohlman, D. Arguijo Mendoza, S. Ding, M. Gao, H. Dressman, I. D. Iliev, S. M.  
639 Lipkin, X. Shen, The cancer microbiome atlas: a pan-cancer comparative analysis to  
640 distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* **29**, 281-  
641 298 e285 (2021).
- 642 24. S. Borchmann, An atlas of the tissue and blood metagenome in cancer reveals novel  
643 links between bacteria, viruses and cancer. *Microbiome* **9**, 94 (2021).
- 644 25. G. D. Poore, E. Kopylova, Q. Zhu, C. Carpenter, S. Fraraccio, S. Wandro, T. Kosciolk,  
645 S. Janssen, J. Metcalf, S. J. Song, J. Kanbar, S. Miller-Montgomery, R. Heaton, R.  
646 McKay, S. P. Patel, A. D. Swafford, R. Knight, Microbiome analyses of blood and  
647 tissues suggest cancer diagnostic approach. *Nature* **579**, 567-574 (2020).
- 648 26. H. S. Cheng, S. P. Tan, D. M. K. Wong, W. L. Y. Koo, S. H. Wong, N. S. Tan, The Blood  
649 Microbiome and Health: Current Evidence, Controversies, and Challenges. *Int J Mol*  
650 *Sci* **24**, (2023).
- 651 27. C. C. S. Tan, K. K. K. Ko, H. Chen, J. Liu, M. Loh, S. G. K. H. Consortium, M. Chia, N.  
652 Nagarajan, No evidence for a common blood microbiome based on a population  
653 study of 9,770 healthy humans. *Nat Microbiol* **8**, 973-985 (2023).
- 654 28. J. Abed, N. Maalouf, A. L. Manson, A. M. Earl, L. Parhi, J. E. M. Emgard, M. Klutstein,  
655 S. Tayeb, G. Almogy, K. A. Atlan, S. Chaushu, E. Israeli, O. Mandelboim, W. S. Garrett,  
656 G. Bachrach, Colon Cancer-Associated Fusobacterium nucleatum May Originate  
657 From the Oral Cavity and Reach Colon Tumors via the Circulatory System. *Front Cell*  
658 *Infect Microbiol* **10**, 400 (2020).
- 659 29. T. N. Y. Kwong, X. Wang, G. Nakatsu, T. C. Chow, T. Tipoe, R. Z. W. Dai, K. K. K. Tsoi,  
660 M. C. S. Wong, G. Tse, M. T. V. Chan, F. K. L. Chan, S. C. Ng, J. C. Y. Wu, W. K. K. Wu, J.  
661 Yu, J. J. Y. Sung, S. H. Wong, Association Between Bacteremia From Specific Microbes  
662 and Subsequent Diagnosis of Colorectal Cancer. *Gastroenterology* **155**, 383-390 e388  
663 (2018).
- 664 30. A. Gihawi, Y. Ge, J. Lu, D. Puiu, A. Xu, C. S. Cooper, D. S. Brewer, M. Perteau, S. L.  
665 Salzberg, Major data analysis errors invalidate cancer microbiome findings. *mBio*,  
666 e0160723 (2023).

- 667 31. A. Gihawi, C. S. Cooper, D. S. Brewer, Caution regarding the specificities of pan-  
668 cancer microbial structure. *Microb Genom* **9**, (2023).
- 669 32. G. D. Sepich-Poore, D. McDonald, E. Kopylova, C. Guccione, Q. Zhu, G. Austin, C.  
670 Carpenter, S. Fraraccio, S. Wandro, T. Kosciolk, S. Janssen, J. L. Metcalf, S. J. Song, J.  
671 Kanbar, S. Miller-Montgomery, R. Heaton, R. McKay, S. P. Patel, A. D. Swafford, T.  
672 Korem, R. Knight, Robustness of cancer microbiome signals over a broad range of  
673 methodological variation. *Oncogene*, (2024).
- 674 33. G. I. Austin, T. Korem, Compositional transformations can reasonably introduce  
675 phenotype-associated values into sparse features. *mSystems* **10**, e0002125 (2025).
- 676 34. Y. Ge, J. Lu, D. Puiu, M. Revsine, S. L. Salzberg, Comprehensive analysis of microbial  
677 content in whole-genome sequencing samples from The Cancer Genome Atlas  
678 project. *Science Translational Medicine* **17**, (2025).
- 679 35. S. J. Salter, M. J. Cox, E. M. Turek, S. T. Calus, W. O. Cookson, M. F. Moffatt, P.  
680 Turner, J. Parkhill, N. J. Loman, A. W. Walker, Reagent and laboratory contamination  
681 can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87  
682 (2014).
- 683 36. E. K. Gregory D. Sepich-Poore, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio,  
684 Stephen Wandro, Tomasz Kosciolk, Stefan Janssen, Jessica Metcalf, Se Jin Song, Jad  
685 Kanbar, Sandrine Miller-Montgomery, Robert Heaton, Rana Mckay, Sandip Pravin  
686 Patel, Austin D Swafford, Rob Knight, Reply to: Caution Regarding the Specificities of  
687 Pan-Cancer Microbial Structure. *BioRxIV*, (2023).
- 688 37. G. R. Thompson, 3rd, J. H. Young, Aspergillus Infections. *N Engl J Med* **385**, 1496-  
689 1509 (2021).
- 690 38. A. F. Pedrosa, C. Lisboa, A. Goncalves Rodrigues, Malassezia infections: a medical  
691 conundrum. *J Am Acad Dermatol* **71**, 170-176 (2014).
- 692 39. M. Parapouli, A. Vasileiadis, A. S. Afendra, E. Hatziloukas, Saccharomyces cerevisiae  
693 and its industrial applications. *AIMS Microbiol* **6**, 1-31 (2020).
- 694 40. J. Vinhal Costa Orsine, R. Vinhal da Costa, M. R. Carvalho Garbi Novaes, Mushrooms  
695 of the genus Agaricus as functional foods. *Nutr Hosp* **27**, 1017-1024 (2012).
- 696 41. C. Shi, S. Liu, X. Tian, X. Wang, P. Gao, A TP53 mutation model for the prediction of  
697 prognosis and therapeutic responses in head and neck squamous cell carcinoma.  
698 *BMC Cancer* **21**, 1035 (2021).
- 699 42. G. Magiorkinis, P. C. Matthews, S. E. Wallace, K. Jeffery, K. Dunbar, R. Tedder, J. L.  
700 Mbisa, B. Hannigan, E. Vayena, P. Simmonds, D. S. Brewer, A. Gihawi, G. Rallapalli, L.  
701 Lahnstein, T. Fowler, C. Patch, F. Maleady-Crowe, A. Lucassen, C. Cooper, Potential  
702 for diagnosis of infectious disease from the 100,000 Genomes Project Metagenomic  
703 Dataset: Recommendations for reporting results. *Wellcome Open Research* **4**,  
704 (2019).
- 705 43. I. N. Olomu, L. C. Pena-Cortes, R. A. Long, A. Vyas, O. Krichevskiy, R. Luellwitz, P.  
706 Singh, M. H. Mulks, Elimination of "kitome" and "splashome" contamination results  
707 in lack of detection of a unique placental microbiome. *BMC Microbiol* **20**, 157 (2020).
- 708 44. C. Aurrecoechea, A. Barreto, E. Y. Basenko, J. Brestelli, B. P. Brunk, S. Cade, K.  
709 Crouch, R. Doherty, D. Falke, S. Fischer, B. Gajria, O. S. Harb, M. Heiges, C. Hertz-  
710 Fowler, S. Hu, J. Iodice, J. C. Kissinger, C. Lawrence, W. Li, D. F. Pinney, J. A. Pulman,  
711 D. S. Roos, A. Shanmugasundram, F. Silva-Franco, S. Steinbiss, C. J. Stoeckert, Jr., D.  
712 Spruill, H. Wang, S. Warrenfeltz, J. Zheng, EuPathDB: the eukaryotic pathogen  
713 genomics database resource. *Nucleic Acids Res* **45**, D581-D591 (2017).

- 714 45. D. H. Parks, M. Chuvochina, C. Rinke, A. J. Mussig, P. A. Chaumeil, P. Hugenholtz,  
715 GTDB: an ongoing census of bacterial and archaeal diversity through a  
716 phylogenetically consistent, rank normalized and complete genome-based  
717 taxonomy. *Nucleic Acids Res* **50**, D785-D794 (2022).
- 718 46. M. Lechner, J. Liu, L. Masterson, T. R. Fenton, HPV-associated oropharyngeal cancer:  
719 epidemiology, molecular biology and clinical management. *Nat Rev Clin Oncol* **19**,  
720 306-327 (2022).
- 721 47. R. Hurst, D. S. Brewer, A. Gihawi, J. Wain, C. S. Cooper, Cancer invasion and  
722 anaerobic bacteria: new insights into mechanisms. *J Med Microbiol* **73**, (2024).
- 723 48. Genomics\_England. (2017).
- 724 49. PCAWG, PCAWG - PanCancer Analysis of Whole Genomes. (2019).
- 725 50. Genomics England, The National Genomic Research Library V5.1. 2020  
726 (10.6084/m9.figshare.4530893/7).
- 727 51. JGI. ([https://archive.jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-  
728 user-guide/](https://archive.jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/)).
- 729 52. R. Eisenhofer, J. J. Minich, C. Marotz, A. Cooper, R. Knight, L. S. Weyrich,  
730 Contamination in Low Microbial Biomass Microbiome Studies: Issues and  
731 Recommendations. *Trends Microbiol* **27**, 105-117 (2019).
- 732 53. C. E. Swenson, R. T. Sadikot, Achromobacter respiratory infections. *Ann Am Thorac  
733 Soc* **12**, 252-258 (2015).
- 734 54. C. Racz, R. Petrovski, C. T. Saunders, I. Chorny, S. Kruglyak, E. H. Margulies, H. Y.  
735 Chuang, M. Kallberg, S. A. Kumar, A. Liao, K. M. Little, M. P. Stromberg, S. W. Tanner,  
736 Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms.  
737 *Bioinformatics* **29**, 2041-2043 (2013).
- 738 55. C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, R. K. Cheetham, Strelka:  
739 accurate somatic small-variant calling from sequenced tumor-normal sample pairs.  
740 *Bioinformatics* **28**, 1811-1817 (2012).
- 741 56. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic Local Alignment  
742 Search Tool. *J Mol Biol* **215**, (1990).
- 743 57. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler  
744 transform. *Bioinformatics* **25**, 1754-1760 (2009).
- 745 58. J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J.  
746 P. Mesirov, Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26 (2011).

747  
748  
749  
750

751 **Figure 1 –Pan Cancer Microbial Structure in Genomics England cohort.** A) Microbial load shown as total  
752 bacterial reads per million human reads across tumour types. B) t-SNE plot of Kraken results of 8,103 non-FFPE,  
753 PCR-free, primary tumour samples within Genomics England’s 100,000 Genomes Project that have been reduced  
754 to include 495 genera (table S3). Each point represents a sample coloured by tumour site. t-SNE was carried out  
755 on a matrix of Spearman’s correlation values between samples. This analysis shows on only the predominant  
756 tumour types in the cohort. Tumour types with smaller sample sizes were omitted: carcinoma of unknown  
757 primary, childhood, endocrine, nasopharyngeal, other, sinonasal, testicular, and upper gastrointestinal. Please  
758 note that tumour types such as hepatopancreatobiliary cancer also contain multiple cancer types.

759  
760  
761  
762  
763

760 **Figure 2 –Performance of machine learning classifiers to predict one tumour type from all others based on  
761 microbial content in Genomics England.** Data used is the raw community matrices data (Voom transformed).  
762 Tumours included are only primary tumours, PCR free from fresh frozen tissue. Carcinoma of unknown primary,  
763 nasopharyngeal, 'other', endocrine and sinonasal tumours have been excluded due to small sample sizes.

764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
  
781

**Figure 3 –Translational opportunities for identifying microbial DNA in human cancer sequencing data.** A) Alphapapillomavirus classification in oral/oropharyngeal primary (triangle) and metastatic (circle) tumour samples. The y-axis denotes the number of genus-level Alphapapillomavirus reads and the x-axis denotes clinical diagnostic test results for HPV. Point color indicates the consequence of small gene variants of the *TP53* gene. Samples with no consequence detected were presumed to be wild type (WT). 38 samples were HPV-negative by clinical diagnostics, and 10 HPV-positive. B) Alignment of HTLV-1-classified reads (Kraken) from breast tumour and germline samples from one participant. The image shows the alignment viewed with IGV. The top track denotes coverage for particular regions (maximum coverage = 13). Coloured regions indicate single nucleotide differences present in the reads and not the reference genomes (orange=G, blue=C, red=T, green=A). In total 172 quality-trimmed, human-depleted reads were subject to alignment (66 and 106 reads from the tumour and germline sample, respectively). C) Kaplan-Meier plot investigating survival in the sarcoma cohort for samples positive for at least one ABBS genus (Anaerobic Bacterial Biomarker Set). This includes *Fenollaria*, *Ezakiella*, *Peptoniphilus*, *Porphyromonas*, *Anaerococcus* and *Fusobacterium*. P=0.0093 was obtained using the log-rank test. Time was measured by years from sample collection.