The Landscape of Microbial Associations in Human Cancer

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

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
6 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
13 Lineberger Comprehensive Cancer Center, University of North 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

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

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

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

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

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

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

Results

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

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

Elucidating Pan-Cancer Microbial Structure

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

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

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

Fungal Genera in Genomics England Dataset

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

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 HLTV-1 was identified as a reportable actionable finding (*42*). All reads in our current analysis uniquely hit HTLV-1 sequences (E-values < $1x10^{-70}$ 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,*

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

Discussion

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

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

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

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

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

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

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

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

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

Materials and Methods

Study design

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

Data

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

Statistical analysis

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

For survival analysis, metadata and clinical data was accessed via Rlabkey API within Genomics England's research environment using release version "main-programme_v12_2021_05_06". Date of death was found in either "mortality" or "death_details" datasets, which are provided to Genomics England from the Office of National Statistics and NHS Digital, respectively. For non-deceased participants, date when they were last seen was inferred from the most recent event from "hes_ae" "hes_apc"

"hes_cc" "hes_op" which detail hospital episode statistics from accident and emergency, admitted patient care, critical care and outpatients respectively. Date of tumour collection was obtained from the cancer_analysis dataset. Days to event was calculated as time from sample collection until date of death or the date the participant was last seen and was divided by 365 to convert to years. Survival objects were created using the Surv function (survival R package, version 3.2.3). Survival models were fit with the survfit function (survival R package) and differences examined using log-rank test. Figures were produced with ggsurvplot function (survminer R package, version 0.4.7). Sarcoma disease subtype was inferred from disease_sub_type of the cancer_analysis data.

Taxonomic Classification of Tumour Whole Genome Sequences

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

Feature Selection and Dimensionality Reduction

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

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

Mutation Calling and Analysis

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

Clinical HPV Diagnostics

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

HTLV-1 Investigation

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

HTLV-1 reference genome (NC_001436.1) using BWA-MEM (*57*) with standard parameters which was subsequently visualised with IGV version 2.9.4 (*58*).

Supplementary Materials:

Figures S1-13
Tables S1-15 (in data file S1)
MDAR checklist
Supplementary Materials and Methods

Acknowledgements

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

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

Author Contributions

AG is responsible for Methodology, Formal Analysis, Investigation, Curated Data, Writing the Original Draft, Review & Editing, Data Visualisation and Project Management.
SDN curated data on sarcoma subtypes.
AG, CSC and DSB Conceptualized the study design and acquired funding for the project and are responsible for the interpretation of study data.
RH, DSB and CC are responsible for supervision and project management.
AGS, LM, ML, TRF TMJ and HMW advised on the analysis an interpretation of study data (Alphapapillomavirus).
AR advised on formal analysis (machine learning classifiers on metadata)
AG, HMW, JC, JOG, RAE, DCW, GMJ, GM, AGS, LM, ML, TRF, TMJ, AF, SDN, AR, RH, CSC, DSB reviewed drafts of the manuscript and helped critique for important intellectual content.

Competing Interests

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

Data and materials availability

All data associated with this study are in the paper or supplementary materials. The Kraken database consisting of GRCh38 and all bacteria, viral (which includes bacteriophages), fungal and protozoal genomes at the scaffold level and above (constituent genomes can be found at https://zenodo.org/records/15739381). Community matrices, analysis scripts and DNA reads unmapped to the human genome are available within the Genomics England research environment for researchers to access. The community matrix used can be located at the file path: /re_gecip/shared_all_GeCIPs/Abe/all_kraken_community.tsv.

References

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

associated with clinicopathological features, specific genomic alterations and better survival. *medRxiv*, (2023).

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

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

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

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

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

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

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.