Gastric Cancer (2025) 28:358–371
https://doi.org/10.1007/s10120-025-01588-9

ORIGINAL ARTICLE

The relationship between the gastric cancer microbiome and clinicopathological factors: a metagenomic investigation from the 100,000 genomes project and The Cancer Genome Atlas

Mary E. Booth[1] · Henry M. Wood[1] · Mark A. Travis[2] · Genomics England Research Consortium[3,4] · Phil Quirke[1] · Heike I. Grabsch[1,5]

Received: 27 November 2024 / Accepted: 15 January 2025 / Published online: 17 February 2025

Abstract

Background Findings from previous gastric cancer microbiome studies have been conflicting, potentially due to patient and/or tumor heterogeneity. The intratumoral gastric cancer microbiome and its relationship with clinicopathological variables have not yet been characterized in detail. We hypothesized that variation in gastric cancer microbial abundance, alpha diversity, and composition is related to clinicopathological characteristics.

Methods Metagenomic analysis of 529 GC samples was performed, including whole exome sequencing data from The Cancer Genome Atlas (TCGA) and whole genome sequencing data from the 100,000 Genomes Project. Microbial abundance, alpha diversity, and composition were compared across patient age, sex, tumor location, geographic origin, pathological depth of invasion, pathological lymph node status, histological phenotype, microsatellite instability status, and TCGA molecular subtype.

Results Gastric cancer microbiomes resembled previous results, with *Prevotella*, *Selenomonas*, *Stomatobaculum*, *Streptococcus*, *Lactobacillus*, and *Lachnospiraceae* commonly seen across both cohorts. Within the TCGA cohort, microbial abundance and alpha diversity were greater in gastric cancers with microsatellite instability, lower pathological depth of invasion, intestinal-type histology, and those originating from Asia. Microsatellite instability status was associated with microbiome composition in both cohorts. Sex and pathological depth of invasion were associated with microbiome composition in the TCGA cohort.

Conclusion The intratumoral gastric cancer microbiome appears to differ according to clinicopathological factors. Certain clinicopathological factors associated with favourable outcomes in gastric cancer were observed to be associated with greater microbial abundance and diversity. This highlights the need for further work to understand the underlying biological mechanisms behind the observed microbiome differences and their potential clinical and therapeutic impact.

Keywords Gastric cancer · Gastrointestinal microbiome · Metagenomics

✉ Heike I. Grabsch
h.grabsch@maastrichtuniversity.nl

[1] Division of Pathology & Data Analytics, Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, UK

[2] Lydia Becker Institute for Immunology and Inflammation, Wellcome Trust Centre for Cell-Matrix Research, Division of Immunology, Immunity to Infection and Respiratory Medicine, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, UK

[3] Genomics England, London, UK

[4] William Harvey Research Institute, Queen Mary University of London, London EC1M 6BQ, UK

[5] Department of Pathology, GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Center+, Maastricht, The Netherlands

Introduction

Gastric cancer (GC) is the fifth most common cancer globally and the fifth most common cause of cancer death worldwide. In 2022, there were over 960,000 new cases and 660,175 deaths attributable to GC [1]. *Helicobacter pylori* (*H. pylori*) infection is known to increase the risk of developing GC in some individuals and is the only bacterium identified as an International Agency for Research on Cancer (IARC) class one carcinogen [2, 3]. Increased availability and reduced cost of microbial sequencing has progressed GC microbiome research beyond focusing on single microorganisms towards analysis of whole microbiomes of GC patient cohorts.

Recent studies suggest a lower intratumoral microbial alpha (within sample) diversity in patients with GC, compared to the mucosa of patients with GC precancerous conditions [4–7] or healthy controls [8]. Comparisons of alpha diversity of matched tumor and adjacent tissue suggested increased diversity in GC relative to adjacent tissue, possibly related to the reduced abundance of *H. pylori* in established GC [9–11]. Several studies have compared the GC microbiome composition to the microbiome of normal mucosa or precursor lesions [4–6, 8, 12–14]. The overlap in taxa found more commonly in GC between studies is relatively small. There is, therefore, currently no consensus dysbiotic microbiome associated with GC. GC is a heterogenous disease, with variation in incidence and patient outcomes according to geographical origin, sex, histology, and molecular phenotype [1, 15–18]. However, the contribution of different patient and tumor characteristics to this variation is not known and microbiome research has largely considered GC as a single disease. A better understanding of the potential relationship between patient- and tumor-specific characteristics and the GC microbiome is needed.

In metagenomic approaches, non-human reads are aligned to microbial databases. This approach allows the analysis of large patient populations from sequencing databases, such as The Cancer Genome Atlas (TCGA) [19]. Whole metagenome analysis has the advantage of enabling higher level taxonomic identification to species or even subspecies level than 16S ribosomal RNA sequencing, and deeper microbial coverage than whole exome sequencing.

One rarely considered factor in microbiome analyses, including but not limited to metagenomic analyses, is contaminating microbial DNA related to the material sampling process or the laboratory environment [20–24]. Contamination may distort results and its effect has been shown to be exaggerated in studies of low microbial biomass [25, 26]. Most GC microbiome studies, so far, have not included an *in silico* decontamination processes; this may have contributed to the previously reported conflicting results.

The relationship between the microbiome and selected clinicopathological factors has been explored in a small number of studies [11, 27–29]. These studies have predominantly focused on patients from Asia and have not included thorough decontamination processes.

We hypothesized that GC microbial abundance, alpha diversity, and composition vary according to clinicopathological characteristics. We aimed to characterize the microbiome of GC and explore relationships between the microbiome and clinicopathological features using sequencing data and accompanying clinicopathological data from GC from the 100,000 Genomes Project and TCGA, incorporating a custom *in silico* decontamination process. Through identification of patient- and tumor-specific factors associated with differences in the GC microbiome, we aimed to better understand the role of the microbiome in this heterogenous disease.

Methods

Genomes Project data acquisition

Whole genome sequencing data of fresh frozen primary gastric adenocarcinoma and matched blood samples, plus clinical metadata from the 100,000 Genomes Project were accessed within the Genomics England Research Environment [30]. All analyses of Genomics England data were performed within the Genomics England Research Environment.

TCGA data acquisition

Exome sequencing data of fresh frozen primary gastric adenocarcinoma and matched blood samples plus virtual slide images from the TCGA stomach adenocarcinoma project were obtained from the National Institute of Health National Cancer Institute Genomic Data Commons Data Portal [31]. Basic clinical characteristics were obtained from the University of California Xena TCGA hub, (https://tcga.xenahubs.net) and Liu *et al.* [32]. TCGA GC molecular subtype data were obtained from the TCGA Research Network [19]. For the majority of cases, Lauren histological classification was publicly available [33]; for cases where Lauren classification was not available, the classification was provided by a gastrointestinal histopathologist after reviewing the slide images. All CIBERSORT [34] immune cellular fraction estimates and immune subtypes were obtained from Thorsson [35]. Estimates of six pre-selected immune cells (lymphocytes, neutrophils, macrophages, dendritic cells, eosinophils, and mast cells) and immune subtypes were analysed in exploratory analyses.

Microsatellite instability (MSI) status and TCGA molecular subtypes

For the TCGA cohort, MSI status was obtained from previously published data [36]. In the present study, MSI-low cases were grouped with microsatellite stable (MSS) cases, since MSS and MSI-low were previously reported to be similar with respect to mutations per Mb (mut/Mb) [37]. The TCGA molecular subtype of the TCGA GC cohort was obtained from the TCGA Research Network classifications [19].

Since MSI status and TCGA molecular subtype data were not available for the 100,000 Genomes Project cohort, MSI status and TCGA molecular subtype were inferred. MSI status was inferred using the number of somatic coding variants (SCV) per sample. Epstein-Barr Virus (EBV) status was determined using sequencing count (virions per human cell), and DNA ploidy was obtained from metadata tables. The TCGA subtype was subsequently inferred, based upon Bass *et al.* [38]. See Online Resource data for flowchart and thresholds used to infer TCGA subtype.

Metagenomic profiling

Microbiome data were generated from sequencing data using the GATK PathSeq algorithm, aligned against the default PathSeq microbial databases [39]. The PathSeq 'score' output was used for microbial sequencing reads, except for the decontamination steps where unambiguously mapping reads were used.

Decontamination

A modified version of the methodology described by Dohlman [26] was used for *in silico* decontamination. Prevalence was defined as at least two unambiguously mapping reads per taxa per sample. For each species, blood prevalence was compared to tissue prevalence in both the 100,000 Genomes Project and TCGA cohorts. One sided Fisher's exact test was performed for each species-specific comparison, using a significance threshold of q<0·05. An include-list was created from species more prevalent in tissue than in blood (q<0·05) in the 100,000 Genomes Project or TCGA, where blood prevalence was <20% of samples in both cohorts. For all species with q values ≥0·05 and <0·4 from the TCGA cohort, the literature was reviewed and species identified as inhabitants of the digestive or respiratory tracts were manually added to the include-list. In addition, EBV was manually added to the include-list. Decontaminated datasets for both cohorts for downstream analysis were created by filtering the genus and species reads to include only species present on the include-list. All downstream analysis used only the decontaminated datasets.

Statistical analyses

Analyses were conducted in R [40] within RStudio [41], using stringr [42], dplyr[43], qvalue [44], and vegan [45] packages. Due to differences in sequencing methodologies and availability of metadata between the two cohorts, analyses were performed in either one or both cohorts depending on data availability and similarities across.

Clinicopathological variables used in analyses included: age, sex, tumor location, geographical origin, pathological depth of invasion (pT), pathological lymph node status (pN), histological phenotype, MSI status, and TCGA molecular subtype.

To evaluate abundance, total microbial count was calculated for each sample. Within the 100,000 Genomes Project cohort, microbial abundance was represented by microbes per human cell and was calculated using the following formula:

$$microbesperhumancell = \frac{(microbialreads \div microbialgenomesize)}{(humanreads \div humangenomesize)},$$

after adjustment of the human genome size for DNA ploidy and tumor cell content. Within the TCGA cohort, microbial abundance represented the sum of microbial sequencing reads. No adjustment for human genome size was made using the TCGA data, since data were derived from exome reads and therefore human reads were not representative of the whole genome due to overrepresentation of exons (relative to intergenic regions).

Shannon index [46] was calculated for each sample as a measure of alpha diversity. Wilcoxon and Kruskal–Wallis tests were applied for comparisons of categorical variables. Spearman's rank correlation coefficient was calculated for correlation analyses. Permutational multivariate analysis of variance (PERMANOVA, Adonis), using Bray-Curtis dissimilarity index, [47] was used to analyze beta diversity in species between subgroups. The PERMANOVA analysis considered clinicopathological variables with over 85% completeness. To avoid overlapping variables, TCGA molecular subtype was not included since it is a composite variable which includes MSI status, which was already included in PERMANOVA analysis. Samples with no taxa were removed prior to PERMANOVA analysis. Variables significant at beta diversity PERMANOVA analysis were included in analysis of differential abundance according to clinicopathological variables. Multivariable Association

Table 1 Clinicopathological characteristics of samples used within this study, from 100,000 Genomes Project and TCGA cohorts

Characteristic	100,000 Genomes Project – gastro-oesophageal cancer total n = 89 n (%)	TCGA – gastric cancer total n = 440 n (%)
Age, median (interquartile range)	69 (62-77)	67 (58-73)
Unknown	0 (0)	5 (1)
Sex		
Male	69 (78)	283 (64)
Female	20 (23)	157 (36)
Tumor location		
Cardia	27 (30)	90 (20)
Non-cardia	44 (49)	280 (64)
Unknown	18 (20)	70 (16)
Geographic origin		
Asia	0 (0)	69 (15)
Not Asia	89 (100)	314 (71)
Unknown	0 (0)	57 (13)
Pathological depth of invasion (pT)		
1	8 (9)	23 (5)
2	11 (12)	93 (21)
3	32 (36)	198 (45)
4	29 (33)	117 (27)
Unknown	9 (10)	9 (2)
Pathological lymph node status (pN)		
0	20 (23)	131 (30)
1	25 (28)	117 (27)
2	22 (25)	86 (20)
3	13 (15)	88 (20)
Unknown	9 (10)	18 (4)
Histological phenotype		
Diffuse	10 (11)	76 (17)
Intestinal	23 (26)	278 (63)
Mixed	2 (2)	21 (4)
Mucinous	0 (0)	19 (4)
Unknown	54 (61)	46 (10)
MSI status		
MSS	81(91)*	308 (70)
MSI	8(9)*	75 (17)
Unknown	0 (0)	57 (13)
TCGA subgroup		
EBV	3 (3)*	26 (6)
MSl	8 (9)*	64 (15)
CIN	35 (39)*	145 (33)
GS	43 (48)*	58 (13)
Unknown	0 (0)*	147 (33)
Immune subtype[†]		
C1	0 (0)	129 (31)
C2	0 (0)	209 (51)
C3	0 (0)	35 (9)
C4	0 (0)	9 (2)
C5	0 (0)	0 (0)
C6	0 (0)	7 (2)
Unknown	89 (100)	51 (12)

CIN, chromosomal instability; *EBV*, Epstein-Barr virus-positive; *GS*, genomically stable; *MSI*, microsatellite instability; *MSS*, microsatellite stabile; *n*, number

[*] Inferred as described in methods

[†] Immune subtype according to Thorsson *et al.* available for TCGA cohort only

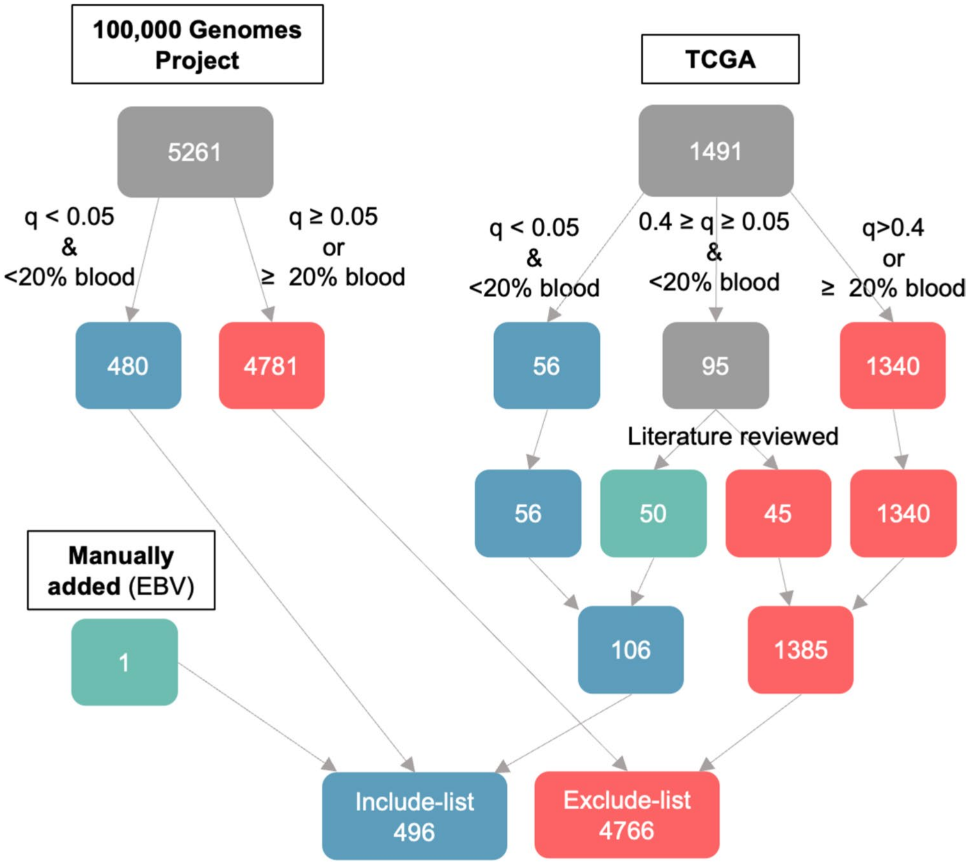


Fig. 1 Overview of the decontamination process. Flow chart to demonstrate categorisation of species into the study include- or exclude-lists, using both 100,000 Genomes Project and TCGA species-level data. The numbers in the final include- and exclude-lists total less than the species upstream, as some species were prevalent within both datasets. *EBV*, Epstein-Barr Virus; *TCGA*, The Cancer Genome Atlas

Discovery in Population-scale Meta-omics Studies (MaAsLin2) [48] was used to perform differential abundance analyses.

Results

Eighty nine tumor samples from patients with gastric or gastro-oesophageal junction adenocarcinoma (GC) were identified in the 100,000 Genomes Project database. All 89 GC samples had whole genome sequencing data for tumor and blood which were used to generate microbial sequencing data. TCGA microbial sequencing data were generated from whole exome sequencing data from 441 tumor samples and 396 matched blood samples from the TCGA GC cohort. Histological review of available slides from the TCGA GC cohort identified one sample as squamous cell carcinoma, therefore, this case was removed from subsequent analysis. The distribution of clinicopathological features within both the 100,000 Genomes Project and TCGA cohorts can be found in Table 1.

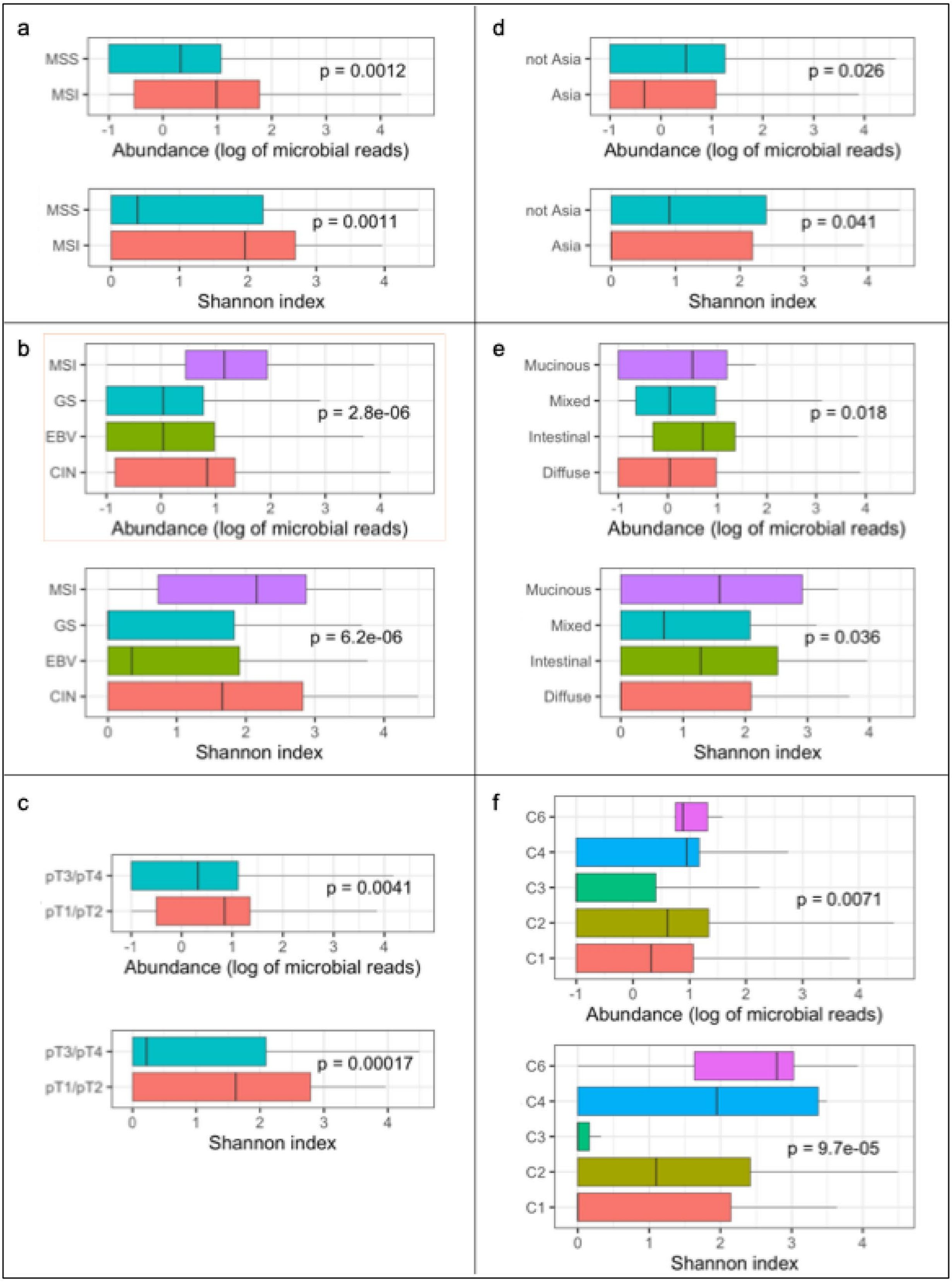
MSI status and TCGA molecular subtypes

MSI status was available for 383 (87%) TCGA GC from previously published data [36]. MSI data were not directly available for the 100,000 Genomes Project data and an inferred MSI status was used for analysis (Online Resource Figure 1): samples where SCV ≥20 mut/Mb were inferred as MSI; samples where SCV <20 mut/Mb were inferred as MSS.

TCGA molecular subtype status was available for 293/440 (67%) GC from the TCGA cohort. Within the 100,000 Genomes Project, thresholds used to determine inferred TCGA molecular subtype were: EBV count 1x10^{-3} per human cell; SCV 20mut/Mb; and DNA ploidy 2·5 (Online Resource Figure 1). The distribution of samples according to inferred MSI and TCGA molecular subtype is shown in Table 1.

Decontamination

Prior to decontamination, 5261 species from 946 genera were present across at least one sample from samples of the 100,000 Genomes Project cohort; 1491 species from


a
MSS
MSI
p = 0.0012
-1 0 1 2 3 4
Abundance (log of microbial reads)
MSS
MSI
p = 0.0011
0 1 2 3 4
Shannon index
d
not Asia
Asia
p = 0.026
-1 0 1 2 3 4
Abundance (log of microbial reads)
not Asia
Asia
p = 0.041
0 1 2 3 4
Shannon index
b
MSI
GS
EBV
CIN
p = 2.8e-06
-1 0 1 2 3 4
Abundance (log of microbial reads)
MSI
GS
EBV
CIN
p = 6.2e-06
0 1 2 3 4
Shannon index
e
Mucinous
Mixed
Intestinal
Diffuse
p = 0.018
-1 0 1 2 3 4
Abundance (log of microbial reads)
Mucinous
Mixed
Intestinal
Diffuse
p = 0.036
0 1 2 3 4
Shannon index
c
pT3/pT4
pT1/pT2
p = 0.0041
-1 0 1 2 3 4
Abundance (log of microbial reads)
pT3/pT4
pT1/pT2
p = 0.00017
0 1 2 3 4
Shannon index
f
C6
C4
C3
C2
C1
p = 0.0071
-1 0 1 2 3 4
Abundance (log of microbial reads)
C6
C4
C3
C2
C1
p = 9.7e-05
0 1 2 3 4
Shannon index

◂**Fig. 2** Boxplots of microbial abundance and Shannon index of TCGA samples according to **a** MSI status *n* = 383, **b** TCGA subtype, *n* = 293, **c** pathological depth of invasion (pT), *n* = 431, **d** geography, *n* = 383, **e** histological subtype, and **f** immune subtype. *CIN*, chromosomal instability; *EBV*, Epstein-Barr virus-positive; *GS*, genomically stable; *MSI*, microsatellite instability; *MSS*, microsatellite stable. *p* values represent Wilcoxon and Kruskal–Wallis statistics for comparisons of two variables and more than two variables, respectively

374 genera were prevalent across at least one sample from the TCGA cohort. 480 species were identified as tissue-resident from the 100,000 Genomes Project cohort and 106 species (comprising of 56 species where q<0·05 and additional 50 species where q<0·4 and identified as inhabitants of digestive or respiratory tracts) were identified as tissue-resident from TCGA cohort. None of the species identified as tissue-resident from only one cohort had a blood prevalence ≥20% in either cohort. EBV was not statistically assigned to either list due to its low prevalence in the 100,000 Genomes Project cohort (4/89 tumor samples; 0/89 blood samples) and absence within the TCGA cohort (0/441 tumor samples; 0/396 blood samples) and was manually added to the include-list. The final include-list consisted of 496 species from 105 genera. Figure 1 illustrates the decontamination process performed to generate the species include-list.

The decontaminated datasets represented 53 and 78% of the original microbial content for the 100,000 Genomes Project and TCGA, respectively. For the TCGA cohort, 135/440 (31%) of samples had no taxa after the decontamination process, whereas all 100,000 Genomes Project samples had taxa remaining after decontamination.

Taxonomic composition

Following decontamination, six genera (*Prevotella, Selenomonas, Stomatobaculum, Streptococcus, Lactobacillus*, and *Lachnospiraceae*) were found to be in the 10 top abundant genera in both cohorts. Two species (*Stomatobaculum longum* and *Lachnospiraceae bacterium oral taxon 082*) were common to the 10 top abundant species within the two cohorts. Many samples from the TCGA cohort had no or very low microbial abundance following the decontamination process. There was considerable variation in microbial composition within both cohorts. Heatmaps for the most abundant genera and species are shown in Online Resource Figures 3 and 4.

Abundance and alpha diversity

Within the 100,000 Genomes Project cohort, no statistically significant relationships were identified between the analysed clinicopathological variables and either microbial abundance or Shannon index. However, when the inferred MSI status of the 100,000 Genomes Project GC cohort was examined in relation to microbial abundance, MSI samples tended to have greater microbial abundance than MSS samples ($p = 0·061$). However, no relationship was identified between MSI status and Shannon index.

In the TCGA cohort, MSI GC was associated with greater microbial abundance ($p = 0·001$) and Shannon index ($p = 0·001$) than MSS GC (Fig. 2a). Within the TCGA cohort, TCGA molecular subtype was associated with microbial abundance ($p < 0·001$) and Shannon index ($p < 0·001$). The boxplots (Fig. 2b) demonstrate highest microbial abundance and Shannon index in the MSI subtype, and lowest abundance and Shannon index in the genomically stable subtype. Within the TCGA cohort, lower pT category (pT1 and pT2) was associated with both, greater microbial abundance ($p = 0·004$), and greater Shannon index ($p < 0·001$) (Fig. 2c). GC from Asia had lower microbial abundance ($p = 0·03$) and lower Shannon index ($p = 0·04$) than samples not from Asia (Fig. 2d). Histological phenotype was associated with microbial abundance ($p = 0·02$) and Shannon index ($p = 0·04$) whereby mucinous GC and intestinal-type GC had greater microbial abundance than diffuse-type GC (Fig. 2e).

Sex, age, tumor location (cardia versus non-cardia), and pN category were not related to microbial abundance or Shannon index in either cohort.

Immune cellular fraction estimates and immune subtypes data generated from CIBERSORT [34] were available for 389/440 (88%) TCGA GC. Estimates of six immune cell types (lymphocytes, neutrophils, macrophages, dendritic cells, eosinophils, and mast cells) were plotted against both total microbial count and Shannon index. No notable relationships were identified between any of the six immune cell types and either abundance or Shannon index. Immune subtype (C1–C6, according to Thorsson et al.) [35], was associated with a statistically significant difference in both microbial abundance ($p = 0·007$) and Shannon index ($p < 0·001$). The boxplots (Fig. 2f) demonstrate highest microbial abundance and Shannon index in samples belonging to the C4 and C6 subtype, although these only represented 4% of analyzed samples.

Beta diversity

The PERMANOVA analyses considered variables with over 85% completeness, in a pre-defined order of: geographical origin (TCGA only), age, sex, histological phenotype (TCGA only), MSI status, pT category, and pN category. Removal of samples with missing variables (both cohorts) and no taxa (TCGA only), resulted in 80 samples from the 100,000 Genomes Project and 240 samples from TCGA, available for PERMANOVA analysis. The analysis indicated

Table 2 Permutational multivariate analysis of variance (PERMANOVA) for intratumoral species within 100,000 Genomes Project and TCGA

	100,000 Genomes Project ($n = 80$)		TCGA ($n = 240$)	
	R^2	*P* value	R^2	*P* value
Geography	–	–	0·00592	0·1246
Sex	0·01415	0·2184	0·00845	0·0237*
Age	0·01383	0·2634	0·00716	0·0559
MSI †	0·01727	0·0360*	0·00740	0·0455*
Histology	–	–	0·01448	0·2042
pT	0·01391	0·386	0·00776	0·0363*
pN	0·01306	0·308	0·00337	0·6058
Residuals	0·92777		0·94546	–

R2, the proportion of the variance in the microbiome explained by each variable; *MSI* microsatellite instability-high versus non high; *pT*, pathological depth of invasion (pT1/pT2 versus pT3/pT4); *pN*, pathological lymph node status (pN0 versus >pN0)

*statistically significant ($p < 0·05$)

†Inferred MSI status for 100,000 Genomes Project cohort

an association between MSI status and microbial composition within both cohorts. In addition, sex and pT category were associated with differences in microbial composition within the TCGA cohort (Table 2).

In the 100,000 Genomes Project cohort, MaAsLin2 analysis detected 12 species and six genera with statistically significant differential abundance between inferred MSI and MSS inferred subtypes. Figure 3 shows such differential abundances at genus level.

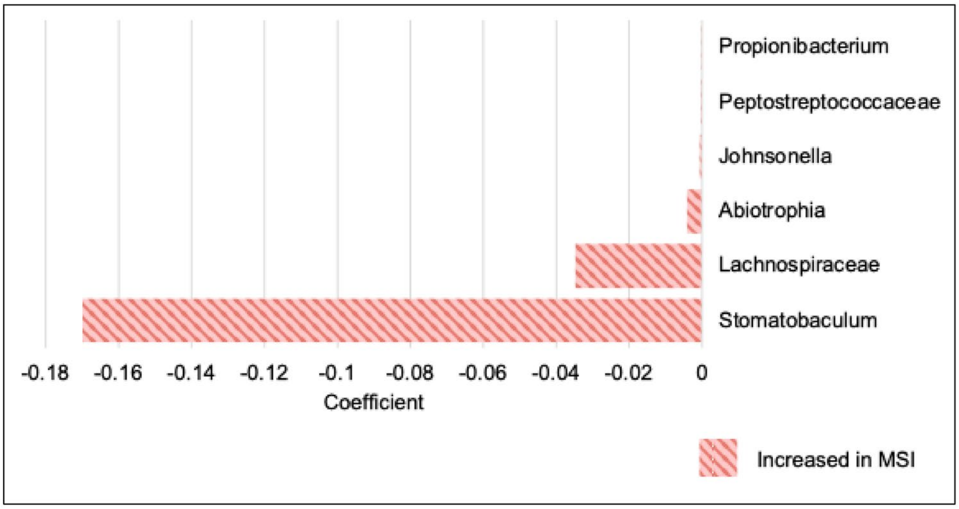
In the TCGA cohort, MaAsLin2 analysis identified 45 species and 12 genera differentially abundant across sex, MSI status and pT category. Differences at genus level are shown in Fig. 4. All differentially abundant taxa according to sex were found more commonly in males than in females. All differentially abundant taxa according to pT category were found more commonly in pT1/pT2 GC compared to pT3/pT4 GC, except for the *Micrococcus* genus and *Micrococcus aloeverae* species, which were found in greater abundance in pT3/pT4 GC. All differentially abundant taxa according to MSI status were found more commonly in MSI GC compared to MSS GC, except for the *Neisseria* genus, which was found in greater abundance in MSS GC. Of note, none of the species or genera identified as differentially abundant according to MSI status within the 100,000 Genomes Project were differentially abundant according to MSI status on multivariate analysis in the TCGA cohort.

Discussion

We explored the relationship between patient- and tumor-specific factors and the GC microbiome. Here, we present the results from an exploratory study of the intratumoral GC microbiome in a total of 529 GC patients, analyzing whole genome sequencing data from the 100,000 Genomes Project and whole exome sequencing data from TCGA. This is the largest GC study to date to use whole genome and whole exome sequencing data to characterize the clinicopathological features associated with the GC microbiome. We identified associations of potential clinical importance through identifying relationships between clinicopathological features and microbial abundance, alpha diversity, and beta diversity.

Within the present study, we further developed the decontamination process as initially described by Dohlman [26], incorporating two separate databases to maximize the number of genuine tumor taxa and minimize contamination. Decontamination is infrequently performed in GC studies; however, the high proportion of the total signal and

Fig. 3 Association of specific genera within the 100,000 Genomes Project samples, according to inferred MSI status by Multivariable Association Discovery in Population-scale Meta-omics Studies (MaAsLin2) ($n = 89$). MaAsLin2 coefficient (effect size) according to MSI status. Red indicates genera enriched in MSI (versus MSS). *MSI*, microsatellite instability; *MSS*, microsatellite stable



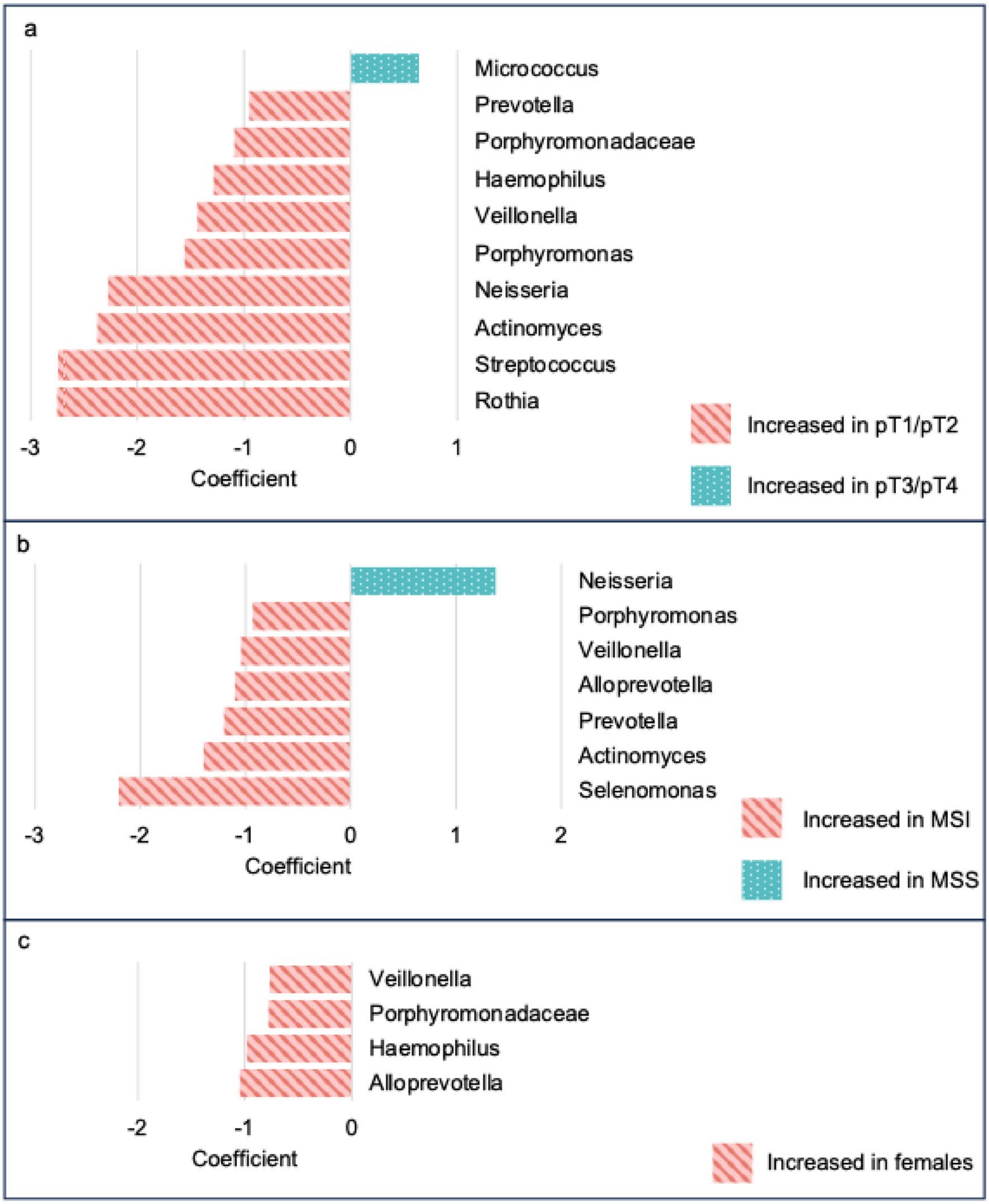


Fig. 4 Association of specific genera within TCGA samples, according to pT category, MSI status, and sex by Multivariable Association Discovery in Population-scale Meta-omics Studies (MaAsLin2) (n = 375). **a** MaAsLin2 coefficient (relative effect size) according to pT category. Red indicates genera enriched in pT1/pT2; green indicates genera enriched in pT3/pT4. **b** MaAsLin2 coefficient according to MSI status. Red indicates genera enriched in MSI; green indicates genera enriched in MSS. **c** MaAsLin2 coefficient according to sex. Red indicates genera enriched in females (versus males). *MSI*, microsatellite instability; *MSS*, microsatellite stable; *pT*, pathological depth of invasion

number of taxa removed emphasize its importance in this setting. This has been previously demonstrated to be particularly important within other low-biomass studies [49, 50]. Furthermore, it has been demonstrated that upstream errors in decontamination processes can have large downstream effect on study results [51]. Eisenhofer *et al*. [49]

and Salter *et al*. [50] suggested an exclude-list approach to decontamination, however, this may result in contaminants being missed (and subsequently being analysed within the dataset). The include-list approach suggested by Dohlman *et al*. was validated by comparing against 16S rRNA amplicon sequencing of matched fresh colorectal cancer tissue, resulting in the absence of putative contaminants; this method has subsequently been used in an analysis of over 2000 colorectal cancers [52]. The include-list approach by Dohlman *et al*. may still result in the exclusion of important taxa, as demonstrated by *Escherichia coli* which did not initially meet the criteria for tissue-resident bacterium in The Cancer Microbiome Atlas analysis, resulting in some further manual curation [26]. These studies collectively guided the two-step decontamination approach taken in the present study of low biomass tissue; first a statistical algorithm was applied to exclude known and unknown contaminants, then borderline taxa were reviewed and added to the include-list where biological evidence was sufficient. This is the first study to systematically analyse the relationship between GC clinicopathological characteristics and the GC microbiome using data from patients across the globe using an *in silico* decontamination process.

We found that microbiome abundance, diversity, and composition differ in GC in relation to MSI status of the tumor. Consistent with our findings in GC, MSI status has also been associated with microbial changes in colorectal cancer [53–57]. A recent study by Byrd *et al.* has also investigated the relationship between MSI status and the microbiome within the TCGA cohort [29]. In line with our findings, this study also found an association between intratumoral microbes and MSI status across stomach, colorectal, and endometrial cancers. The methodology used to generate the microbiome abundance data from whole transcriptome and whole genome sequencing data in this study has since been identified as flawed, although to what extent that affects the validity of this paper by Byrd *et al.* is unclear [51, 58]. Whilst the extent to which the relationship between MSI status and the microbiome is associative or causative is uncertain, it is possible that the local tumor environment of MSI GC somehow facilitates a more abundant, more diverse microbiome.

In addition to MSI status, our results from the TCGA analysis suggest greater microbial abundance and alpha diversity in lower pT category GC, as well as a difference in microbial composition between pT1/pT2 and pT3/pT4 GC. These findings may reflect a difference in host environment in that fewer taxa, or only certain specific taxa, are able to invade beyond the muscle wall. These findings are consistent with those of a study investigating only Asian patients [27], where the microbial composition differed according to stage. This study did not identify alpha diversity differences according to stage, which may indicate that changes are more related to the pT classification than overall staging, which incorporates pT and pN classification, as well as the presence or absence of distant metastasis. Molecular differences according to tumor depth have previously been reported in GC [59, 60]; thus, when considering this in the context of our findings, one could speculate that spatial differences in tumor microenvironment influence the local microbiome. Alternatively, a more abundant, more diverse microbiome may be protective against greater tumor invasion, potentially through increased local immune surveillance.

The increased microbial abundance and alpha diversity seen in intestinal-type GC compared to diffuse-type GC may result from differences in the tumor microenvironment. In line with our findings, lower alpha diversity was observed in patients with diffuse-type GC relative to intestinal-type GC, in a study of 64 GC samples from Lithuanian patients [28]. Previous studies in GC have demonstrated evidence of molecular differences [19, 33], as well as differences in the ratio of tumor to stroma [61], according to histological phenotype. Existing data from a single-cell RNA sequencing analysis of GC is suggestive of differences in the proportions of plasma cells and KLF2-expressing epithelial cells between diffuse- and intestinal-type GC [62]. It is possible that molecular variability between diffuse-type and intestinal-type GC could result in differences in microbial abundance and alpha diversity.

We decided to further explore a potential relationship between the GC microbiome and immune cells in the tumor microenvironment, using CIBERSORT data from TCGA, to try to further understand the observed relationships between the microbiome and clinicopathological characteristics including MSI status, pT category, and histological phenotype. The immune subtypes associated with greater microbial abundance and Shannon diversity (C4 and C6) represented the lymphocyte depleted and TGF-b dominant subtypes, respectively, although only representing 4% of analyzed samples. Our investigation of immune cells with microbial abundance and alpha diversity was not able to provide further insight on this, possibly due to the reliance on whole tumor immune data, which was not spatially orientated. Further investigation of the relationship between the GC microbiome and local immune cells is warranted and spatial techniques should be considered in such investigations.

Notably, MSI (versus MSS) GC, low (versus high) pT category, and intestinal (versus diffuse) histological phenotype—all of which have been observed to have greater microbial diversity and abundance within the present study—are generally recognized to be good prognostic factors in GC [63–66]. In general, higher microbial diversity is thought to be associated with improved health outcomes, including in [67], but not limited to [68], patients with cancer. Considered in this context, our own observations of increased microbial diversity in such "good prognosis" subgroups (MSI, low pT and intestinal histology) warrants further investigation of the role of microbial abundance and diversity in GC behaviour—in particular, whether and through what mechanisms increased microbial diversity may contribute to the improved outcomes observed in these good prognosis groups. If results from future studies support the hypothesis that greater GC microbial abundance and diversity results in superior outcomes, this may inform development of therapeutic investigations to increase microbial abundance and diversity, with the aim of improving GC outcomes.

Our study had some limitations. The 100,000 Genomes Project Cohort had small sample numbers (89) whilst the TCGA cohort was based on exome data, which has a limited ability to facilitate detection of microbiome data in the original sample. Recent work [51] has demonstrated that microbiome data from low biomass exome data can be challenging and easy to over-interpret. While gastric samples do not have the extremely low (or absent) non-human content of other tissue types, care should be taken in analysing samples such as the TCGA cohort. In addition, the geographical and racial composition of the two cohorts differed, with the 100,000 Genomes Project including only individuals from England and TCGA including individuals from across the world, including at least 15% from Asia. Whilst no association was detected within the present study between geography and microbial abundance, alpha diversity, or composition, geography may still account for some of the differences in microbiome composition between the two cohorts. *H. pylori* prevalence varies considerably with geographic location, with higher rates of infection in Asian countries than in Western Europe and the United States [69]. *H. pylori* may influence the gastric mucosal microbiome in both neoplastic [70–72] and non-neoplastic [70, 72, 73] stomachs. Therefore, it is possible that the differences in microbial composition detected between cohorts are related to the geographical heterogeneity that may not have been captured by PERMANOVA analysis due to the over-simplification of geography (i.e., Asia versus not-Asia). Furthermore, the availability of metadata differed between the two cohorts, limiting the investigation to a purely exploratory analysis.

In conclusion, using two separate sequencing databases, we identified microbiome differences in relation to depth of tumor invasion, histological phenotype, and molecular characteristics such as presence of MSI. Our findings further reinforce the notion that the relationship between the GC microbiome and clinicopathological variables is multifactorial and, as such, should be considered when planning, conducting, and interpreting the results from investigational clinical studies. Future work should focus on 1) further functional studies to increase the level of understanding regarding how the local microenvironment may affect and be affected by the tumor microbiome, 2) how the microbiome may affect GC phenotype, and 3) whether ultimately microbiome manipulation could affect outcomes of patients with GC and/or response to specific therapeutic interventions. This study collated 529 GC microbiomes; using this large sample size, we were able to consolidate the findings of previous studies [26, 74, 75] and conduct an evaluation of the relationship between microbial differences and patient and tumor characteristics. The findings of this study underline the potential clinical importance of the GC microbiome and provide a strong rationale for further investigations to improve the depth of understanding in this area.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10120-025-01588-9.

Acknowledgements Genomics England Author list: Ambrose, J. C.; Arumugam, P.; Bevers, R.; Bleda, M.; Boardman-Pretty, F.; Boustred, C. R.; Brittain, H; Brown, M.A.; Caulfield, M. J.; Chan, G. C.; Giess A.; Griffin, J. N.; Hamblin, A.; Henderson, S.; Hubbard, T. J. P.; Jackson, R.; Jones, L. J.; Kasperaviciute, D.; Kayikci, M.; Kousathanas, A.; Lahnstein, L.; Lakey, A.; Leigh, S. E. A.; Leong, I. U. S.; Lopez, F. J.; Maleady-Crowe, F.; McEntagart, M.; Minneci F.; Mitchell, J.; Moutsianas, L.; Mueller, M.; Murugaesu, N; Need, A. C.; O'Donovan P.; Odhams, C. A.; Patch, C.; Perez-Gil, D.; Pereira, M. B.; Pullinger, J.; Rahim, T.; Rendon, A.; Rogers, T.; Savage, K.; Sawant, K; Scott, R. H.; Siddiq, A.; Sieghart, A.; Smith, S. C.; Sosinsky, A.; Stuckey, A.; Tanguy M.; Taylor Tavares, A. L.; Thomas, E. R. A.; Thompson, S. R.; Tucci, A.; Welland, M. J.; Williams, E.; Witkowska, K.; Wood, S. M.; Zarowiecki, M. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project 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. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

Author contributions Conceptualization: M.E.B., H.M.W., M.A.T., P.Q., and H.I.G; Methodology: M.E.B and H.M.W. Formal analysis: M.E.B. and H.M.W.; Investigation: M.E.B. and H.M.W.; Resources: G.E.R.C.; M.E.B., H.M.W., G.E.R.C., and H.I.G., Writing – Original Draft: M.E.B.; Writing – Review & Editing: M.E.B., H.M.W, G.E.R.C., M.A.T., P.Q., and H.I.G; Visualisation: M.E.B.; Funding Acquisition: M.E.B.

Funding This work was supported by a Wellcome Trust 4Ward North PhD fellowship (MB). HG, HW and PQ are supported by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (grant number NIHR203331). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. PQ and HW are also supported by Cancer Research UK Grand Challenge OPTIMISTICC (C10674/A27140). PQ holds a National Institute of Health Research Senior Investigators award. MT is supported by the Medical Research Council, the Kenneth Rainin Foundation, the National Institute for Health and Care Research. The Wellcome Trust Centre for Cell-Matrix Research, University of Manchester, is supported by core funding from the Wellcome Trust [grant number 203128/A/16/Z].

References

1. Ferlay J, Ervik M, Lam F, Laversanne M, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F. Global cancer observatory: cancer today. International Agency for Research on Cancer. Lyon, France. https://gco.iarc.who.int/today. Accessed 4 Feb 2025.
2. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Schistosomes, Liver Flukes and Helicobacter pylori 61:1–241. (1994). PMID: 7715068; PMCID: PMC7681621.
3. Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, et al. Special Report: Policy A review of human carcinogens-Part B: biological agents. Lancet Oncol. 2012;10:321–2. https://doi.org/10.1016/S1470-2045(09)70096-8.
4. Coker OO, Dai Z, Nie Y, Zhao G, Cao L, Nakatsu G, et al. Mucosal microbiome dysbiosis in gastric carcinogenesis. Gut. 2018;67:1024–32. https://doi.org/10.1136/gutjnl-2017-314281.
5. Aviles-Jimenez F, Vazquez-Jimenez F, Medrano-Guzman R, Mantilla A, Torres J. Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer. Sci Rep. 2014;4:1–11. https://doi.org/10.1038/srep04202.
6. Ferreira RM, Pereira-Marques J, Pinto-Ribeiro I, Costa JL, Carneiro F, MacHado JC, et al. Gastric microbial community profiling reveals a dysbiotic cancer-associated microbiota. Gut. 2018;67:226–36. https://doi.org/10.1136/gutjnl-2017-314205.
7. Appiah EM, Yakubu B, Salifu SP. Comprehensive microbial network analysis of gastric microbiome reveal key species affecting gastric carcinogenesis. The Microbe. 2023;1:100009. https://doi.org/10.1016/J.MICROB.2023.100009.
8. Gunathilake MN, Lee J, Choi IJ, Kim YIL, Ahn Y, Park C, et al. Association between the relative abundance of gastric microbiota and the risk of gastric cancer: a case-control study. Sci Rep. 2019;9:1–11. https://doi.org/10.1038/s41598-019-50054-x.
9. Liu X, Shao L, Liu X, Ji F, Mei Y, Cheng Y, et al. Alterations of gastric mucosal microbiota across different stomach microhabitats in a cohort of 276 patients with gastric cancer. EBioMedicine. 2019;40:336–48. https://doi.org/10.1016/j.ebiom.2018.12.034.
10. Chen XH, Wang A, Chu AN, Gong YH, Yuan Y. Mucosa-associated microbiota in gastric cancer tissues compared with non-cancer tissues. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2019.01261.
11. Wang G, Wang H, Ji X, Wang T, Zhang Y, Jiang W, et al. Intratumoral microbiome is associated with gastric cancer prognosis and therapy efficacy. Gut Microbes. 2024. https://doi.org/10.1080/19490976.2024.2369336.
12. Wang L, Xin Y, Zhou J, Tian Z, Liu C, Yu X, et al. Gastric Mucosa-Associated Microbial Signatures of Early Gastric Cancer. Front Microbiol. 2020;11:1548. https://doi.org/10.3389/fmicb.2020.01548.
13. Wang L, Zhou J, Xin Y, Geng C, Tian Z, Yu X, et al. Bacterial overgrowth and diversification of microbiota in gastric cancer. Eur J Gastroenterol Hepatol. 2016;28:261–6. https://doi.org/10.1097/MEG.0000000000000542.
14. Castaño-Rodríguez N, Goh KL, Fock KM, Mitchell HM. Kaakoush NO. Dysbiosis of the microbiome in gastric carcinogenesis. Sci Rep. 2017. https://doi.org/10.1038/s41598-017-16289-2.
15. Cristescu R, Lee J, Nebozhyn M, Kim K-M, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat Med. 2015;21:449–56. https://doi.org/10.1038/nm.3850.
16. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. The Lancet. 2018;391:1023–75. https://doi.org/10.1016/S0140-6736(17)33326-3.
17. de Martel C, Georges D, Bray F, Ferlay J, Clifford GM. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health. 2020;8:e180-90. https://doi.org/10.1016/S2214-109X(19)30488-7.
18. WHO Classification of Tumors Editorial Board. WHO Classification of Tumors, 5th edition: Digestive System Tumors. 5th ed. France: IARC; 2019. ISBN-13: 978-92-832-4499-8
19. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9. https://doi.org/10.1038/nature13480.
20. Kulakov LA, McAlister MB, Ogden KL, Larkin MJ, O'Hanlon JF. Analysis of bacteria contaminating ultrapure water in industrial systems. Appl Environ Microbiol. 2002;68:1548–55.
21. Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. Mol Cell Probes. 2006;20:147–53. https://doi.org/10.1016/j.mcp.2005.09.007.
22. Newsome T, Li B-J, Zou N, Lo S-C. Presence of Bacterial Phage-Like DNA Sequences in Commercial Taq DNA Polymerase Reagents. J Clin Microbiol. 2004;42:2264–7. https://doi.org/10.1128/JCM.42.5.2264-2267.2004.
23. Motley T, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW. Improved Multiple Displacement Amplification (iMDA) and Ultraclean Reagents. BMC Genom. 2014;15:1.
24. Mohammadi T, Reesink HW, Vandenbroucke-Grauls CMJE, Savelkoul PHM. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. J Microbiol Method. 2004. https://doi.org/10.1016/j.mimet.2004.11.018.
25. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016. https://doi.org/10.1186/s13099-016-0103-7.
26. Dohlman AB, Arguijo Mendoza D, Ding S, Gao M, Dressman H, Iliev ID, et al. The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. Cell Host Microbe. 2021;29:281-298.e5. https://doi.org/10.1016/J.CHOM.2020.12.001.

27. Ai B, Mei Y, Liang D, Wang T, Cai H, Yu D. Uncovering the special microbiota associated with occurrence and progression of gastric cancer by using RNA-sequencing. Sci Reports. 2023;13:1–11. https://doi.org/10.1038/s41598-023-32809-9.
28. Lehr K, Nikitina D, Vilchez-Vargas R, Steponaitiene R, Thon C, Skieceviciene J, et al. Microbial composition of tumorous and adjacent gastric tissue is associated with prognosis of gastric cancer. Sci Reports. 2023;13:1–11. https://doi.org/10.1038/s41598-023-31740-3.
29. Byrd DA, Fan W, Greathouse KL, Wu MC, Xie H, Wang X. The intratumor microbiome is associated with microsatellite instability Brief Communication. JNCI: J Nat Cancer Inst. 2023;115:989–93. https://doi.org/10.1093/jnci/djad083.
30. Genomics England. The National Genomics Research and Healthcare Knowledgebase v5 2019.
31. NIH National Cancer Institute. Genomic Data Commons Data Portal n.d. https://portal.gdc.cancer.gov/ (accessed June 5, 2021).
32. Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, et al. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. Cancer Cell. 2018;33:721–35. https://doi.org/10.1016/j.ccell.2018.03.010.
33. Hewitt LC, Saito Y, Wang T, Matsuda Y, Oosting J, Silva S, AN, et al. KRAS status is related to histological phenotype in gastric cancer results from a large multicentre study. Gastric Cancer. 2019;22:1193–203. https://doi.org/10.1007/s10120-019-00972-6.
34. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles HHS Public Access. Nat Methods. 2015;12:453–7. https://doi.org/10.1038/nmeth.3337.
35. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The Immune Landscape of Cancer. Immunity. 2018;48:812-830.e14. https://doi.org/10.1016/j.immuni.2018.03.023.
36. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330. https://doi.org/10.1038/NATURE11252.
37. Ni Huang M, McPherson JR, Cutcutache I, Teh BT, Tan P, Rozen SG. MSIseq: Software for Assessing Microsatellite Instability from Catalogs of Somatic Mutations. Sci Reports. 2015;5:1–10. https://doi.org/10.1038/srep13321.
38. Bass AJ, Thorsson V, Shmulevich I, Reynolds SM, Miller M, Bernard B, et al. Comprehensive Molecular Characterization of Gastric Adenocarcinoma Supplementary Materials. Nature. 2014;1:1.
39. Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, et al. GATK PathSeq: A customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics. 2018;34:4287–9. https://doi.org/10.1093/bioinformatics/bty501.
40. R Core Team. R: A Language and Environment for Statistical Computing 2020. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
41. RStudio Team. RStudio: Integrated Development Environment for R 2021. http://www.rstudio.com/.
42. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations 2019. https://CRAN.R-project.org/package=stringr.
43. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.8. 2022.
44. Storey JD, Bass AJ, Dabney A, Robinson D. qvalue: Q-value estimation for false discovery rate control 2021. R package version 2.32.0. https://doi.org/10.18129/B9.bioc.qvalue.
45. Simpson JO and FGB and MF and RK and PL and DM and PRM and RBO and GL and PS and MHHS and ES and HW. vegan: Community Ecology Package 2020. R package version 2.6-4. https://CRAN.R-project.org/package=vegan.
46. Shannon CE, Weaver W. The Mathematical Theory of Communication. Urbana: The University of Illinois Press; 1949.
47. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol Monogr. 1957;27:325–49. https://doi.org/10.2307/1942268.
48. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable Association Discovery in Population-scale Meta-omics Studies. PLoS Comput Biol. 2021;17:e1009442. https://doi.org/10.1101/2021.01.20.427420.
49. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends Microbiol. 2019;27:105–17. https://doi.org/10.1016/J.TIM.2018.11.003.
50. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:1–12. https://doi.org/10.1186/S12915-014-0087-Z/FIGURES/4.
51. Gihawi A, Ge Y, Lu J, Puiu D, Xu A, Cooper CS, et al. Major data analysis errors invalidate cancer microbiome findings. MBio. 2023;14:1. https://doi.org/10.1128/MBIO.01607-23/SUPPL_FILE/MBIO.01607-23-S0007.XLSX.
52. Cornish AJ, Gruber AJ, Kinnersley B, Chubb D, Frangou A, Caravagna G, et al. The genomic landscape of 2,023 colorectal cancers. Nature. 2024;633:127–36. https://doi.org/10.1038/s41586-024-07747-9.
53. Tahara T, Yamamoto E, Suzuki H, Maruyama R, Chung W, Garriga J, et al. Clinical Studies Fusobacterium in Colonic Flora and Molecular Features of Colorectal Carcinoma. Cancer Res. 2014;74:1311–8. https://doi.org/10.1158/0008-5472.CAN-13-1865.
54. Nosho K, Sukawa Y, Adachi Y, Ito M, Mitsuhashi K, Kurihara H, et al. Association of Fusobacterium nucleatum with immunity and molecular alterations in colorectal cancer. World J Gastroenterol. 2016;22:557–66. https://doi.org/10.3748/wjg.v22.i2.557.
55. Mima K, Nishihara R, Qian ZR, Cao Y, Sukawa Y, Nowak JA, et al. Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. Gut. 2016;65:1973. https://doi.org/10.1136/GUTJNL-2015-310101.
56. Hale VL, Jeraldo P, Chen J, Mundy M, Yao J, Priya S, et al. Distinct microbes, metabolites, and ecologies define the microbiome in deficient and proficient mismatch repair colorectal cancers. Genome Med. 2018;10:1–13. https://doi.org/10.1186/S13073-018-0586-6/FIGURES/4.
57. Vanderbilt CM, Keshinro A, Chen C-T, Babady E, Berger MF, Zehir A, et al. Unique tumor microbiome in microsatellite instability high (MSI-H) colon carcinoma. Cancer Res. 2020;80:6095–6095.
58. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Retraction Note: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature. 2024;631:694–694. https://doi.org/10.1038/s41586-024-07656-x.
59. Naruke A, Azuma M, Takeuchi A, Chikatoshi Katada I, Sasaki T, et al. Comparison of site-specific gene expression levels in primary tumors and synchronous lymph node metastases in advanced gastric cancer. Gastric Cancer. 2015;18:262–70. https://doi.org/10.1007/s10120-014-0357-z.
60. Sundar R, Liu DH, Hutchins GGG, Slaney HL, Silva AN, Oosting J, et al. Spatial profiling of gastric cancer patient-matched primary and locoregional metastases reveals principles of tumor dissemination. Gut. 2020. https://doi.org/10.1136/gutjnl-2020-320805.
61. Aoyama T, Hutchins G, Arai T, Sakamaki K, Miyagi Y, Tsuburaya A, et al. Identification of a high-risk subtype of intestinal-type Japanese gastric cancer by quantitative measurement of the luminal tumor proportion. Cancer Med. 2018;7:4914–23. https://doi.org/10.1002/cam4.1744.

62. Kumar V, Ramnarayanan K, Sundar R, Padmanabhan N, Srivastava S, Koiwa M, et al. Single-Cell Atlas of Lineage States, Tumor Microenvironment, and Subtype-Specific Expression Programs in Gastric Cancer. Cancer Discov. 2022;12:670–91.
63. Choi YY, Bae JM, An JY, Kwon IG, Cho I, Shin HB, et al. Is microsatellite instability a prognostic marker in gastric cancer?: A systematic review with meta-analysis. J Surg Oncol. 2014;110:129–35. https://doi.org/10.1002/JSO.23618.
64. CS Group, Siewert JR, Bottcher K, Stein HJ, Roder JD, Gastric G, et al. Relevant prognostic factors in gastric cancer ten-year results of the German Gastric Cancer Study. Ann Surg. 1998;228:449.
65. Matsuo K, Lee SW, Tanaka R, Imai Y, Honda K, Taniguchi K, et al. T stage and venous invasion are crucial prognostic factors for long-term survival of patients with remnant gastric cancer: a cohort study. World J Surg Oncol. 2021;19:1–7. https://doi.org/10.1186/S12957-021-02400-5/TABLES/3.
66. Petrelli F, Berenato R, Turati L, Mennitto A, Steccanella F, Caporale M, et al. Prognostic value of diffuse versus intestinal histotype in patients with gastric cancer: a systematic review and meta-analysis. J Gastrointest Oncol. 2017;8:148. https://doi.org/10.21037/JGO.2017.01.10.
67. Riquelme E, Zhang Y, Zhang L, Montiel M, Zoltan M, Dong W, et al. Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. Cell. 2019;178:795-806.e12. https://doi.org/10.1016/j.cell.2019.07.008.
68. Claesson MJ, Jeffery IB, Conde S, Power SE, O'connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature. 2012;488:178–84. https://doi.org/10.1038/nature11319.
69. Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, Tanyingoh D, et al. Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. Gastroenterology. 2017;153:420–9. https://doi.org/10.1053/j.gastro.2017.04.022.
70. Li TH, Qin Y, Sham PC, Lau KS, Chu KM, Leung WK. Alterations in Gastric Microbiota after H Pylori Eradication and in Different Histological Stages of Gastric Carcinogenesis. Sci Rep. 2017;7:1–8. https://doi.org/10.1038/srep44935.
71. Nikitina D, Lehr K, Vilchez-Vargas R, Jonaitis LV, Urba M, Kupcinskas J, et al. Comparison of genomic and transcriptional microbiome analysis in gastric cancer patients and healthy individuals. World J Gastroenterol. 2023;29:1202–18. https://doi.org/10.3748/wjg.v29.i7.1202.
72. Xiao W, Ma Z. Influences of Helicobacter pylori infection on diversity, heterogeneity, and composition of human gastric microbiomes across stages of gastric cancer development. Helicobacter. 2022;27:e12899. https://doi.org/10.1111/HEL.12899.
73. Ren R, Wang Z, Sun H, Gao X, Sun G, Peng L, et al. The gastric mucosal-associated microbiome in patients with gastric polyposis. Sci Rep. 2018. https://doi.org/10.1038/s41598-018-31738-2.
74. Wang J, Wang Y, Li Z, Gao X, Huang D. Global Analysis of Microbiota Signatures in Four Major Types of Gastrointestinal Cancer. Front Oncol. 2021. https://doi.org/10.3389/fonc.2021.685641.
75. Rodriguez RM, Hernandez BY, Menor M, Deng Y, Khadka VS. The landscape of bacterial presence in tumor and adjacent normal tissue across 9 major cancer types using TCGA exome sequencing. Comput Struct Biotechnol J. 2020;18:631–41. https://doi.org/10.1016/j.csbj.2020.03.003.