scientific reports

Check for updates

OPEN

Characterization of intra-tumoral microbiota from transcriptomic sequencing of Asian breast cancer

Li-Fang Yeo[1,2], Audrey Weng Yan Lee[1], Phoebe Yon Ern Tee[1], Joyce Seow Fong Chin[1], Bernard K. B. Lee[3], Joanna Lim[1], Soo-Hwang Teo[1] & Jia-Wern Pan[1]✉

The human microbiome has garnered significant interest in recent years as an important driver of human health and disease. Likewise, it has been suggested that the intra-tumoral microbiome may be associated with specific features of cancer such as tumour progression and metastasis. However, additional research is needed to validate these findings in diverse populations. In this study, we characterized the intra-tumoral microbiota of 883 Malaysian breast cancer patients using transcriptomic data from bulk tumours and investigated their association with clinical variables and immune scores. We found that the tumour microbiome was not associated with breast cancer molecular subtype, cancer stage, tumour grade, or patient age, but was weakly associated with immune scores. We also found that the tumour microbiome was associated with immune scores in our cohort using random forest models, suggesting the possibility of an interaction between the tumour microbiome and the tumour immune microenvironment in Asian breast cancer.

Breast cancer is the most common cancer in women across the majority of countries worldwide. Differences in distribution of genetic[1], lifestyle[2] and reproductive factors[3] influence the clinical presentation of breast cancer in different populations. For example, there is a higher prevalence of triple negative breast cancer in women of African descent[4], and a higher prevalence of immune enriched breast cancers in women of Asian descent[5]. Whilst part of these differences may be attributable to differences in population genetics, a large proportion of these differences remain unexplained.

One factor that may potentially explain some of these differences is the microbial community found on and in the human body, also known as the human microbiome. With the advent of next-generation sequencing and decreasing cost to sequence genomes, it has become possible to study the human microbiome in much greater detail. Early studies were mostly focused on characterising the human microbiome[6], but several recent studies have studied their association with cancer and other diseases. Routy et al.[7] reported a retrospective cohort study where cancer patients on antibiotics had shorter progression free-survival and overall survival. Restoring gut microbial diversity via live-bacteria supplements[7] or faecal microbiota transplant[8] improved response to anti-PD1 therapy, suggesting that the gut microbiome may play an important role in treatment outcomes.

Researchers have also been interested in the intra-tumoral microbiome and its association with cancer, though this has been more challenging to study due to its low biomass and accessibility. Recently, intra-tumoral bacteria were found to mostly reside within cancer or immune cells, with each tumour type shown to have a distinct microbiota composition[9]. This landmark paper also showed that breast tumours had the richest and most diverse intra-tumoral microbiome, which was associated with clinical subtypes. Other recent studies have demonstrated the ability of intra-tumoral bacteria to induce the migration of cancer cells and promote cancer progression[10] and metastasis[11].

Notably, global studies that compare the microbiome of different ethnic groups suggest that population-specific tumour microbiomes may exist. For example, a recent study reported differences between tumoral microbiota composition between Caucasians and African Americans but observed no significant differences for Asians[12]. This mirrored findings from Parida et al.[13], who also found differences between Caucasians and African Americans but not Asians. However, both papers included only a very small number of Asian patients in their analyses. This demonstrates the lack of Asian representation in global cancer microbiome studies, which may in turn lead to false assumptions regarding the generalizability of microbiome studies to the wider Asian population.

[1]Cancer Research Malaysia, Subang Jaya, Malaysia. [2]Department of Internal Medicine, University of Turku, Turku, Finland. [3]Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia. ✉email: jiawern.pan@cancerresearch.my

In this study, we characterized the tumoral microbiota of 883 Malaysian breast cancer patients using transcriptomic data from bulk tumours and investigated their association with clinical variables and immune scores. We found that the tumour microbiome was not associated with breast cancer molecular subtype, cancer stage, tumour grade, or patient age, but was weakly associated with immune scores. We also found that the tumour microbiome was associated with immune scores in our cohort using random forest models, suggesting the possibility of interactions between the tumour microbiome and the tumour immune microenvironment in Asian breast cancer.

Methods

Biospecimen collection and data generation

RNA-seq data that was generated by Pan et al.[5] and Pan et al.[14] were used to discover the presence of microbes in fresh frozen tumours from 977 breast cancer patients from the Malaysian Breast Cancer (MyBrCa) cohort recruited at Subang Jaya Medical Centre ($n = 843$) and University Malaya Specialist Centre ($n = 134$), Malaysia. As the sequencing was conducted in two separate batches, the earlier batch was used as a discovery cohort ($n = 558$), and the latter as a validation cohort ($n = 419$) (Supplementary Fig. 1). Immune scores included in the analysis were scored as described in Pan et al.[5]; in brief, scoring was done via gene set variation analyses (GSVA) of different immune gene sets retrieved from literature, as cited in turn in our results.

Data quality assessment and read alignment

RNA-seq reads that mapped to hs38r42 human genome using STAR aligner were removed[15]. Non-human, unmapped reads were retained and mapped to the Kraken2 32GB database[16]. Relative abundance of microbial reads from Kraken2 were estimated using Bracken[17]. Read count tables were created for each taxonomic level by using the kreport2mpa.py script from KrakenTools[18].

Alpha and beta diversity analyses

Reads were converted to relative abundance. Intra-group (alpha) diversity was determined using the number of observed species and Shannon index. Inter-group (beta) diversity was measured using a Bray–Curtis dissimilarity matrix, plotted using unsupervised, multi-dimensional scaling (MDS) method and visualized on a PCoA (Principal Coordinates Analysis) plot. Shepherd's stress test was used to measure goodness-of-fit of the model, that is how well the reduced dimensions reflect the original dissimilarity structure. Beta diversity was also measured using supervised ordination, dbRDA (distance-based Redundancy Analysis). PERMANOVA (Permutational Multivariate ANOVA) was used to calculate the differences between groups controlled by covariates. Covariates included in the analysis were PAM50 subtype, age at diagnosis, cancer stage, tumour grade, treatment used, and ethnicity. PERMDISP (Permutational Multivariate Analysis of Dispersion) was calculated to ensure homogenous dispersion was observed in the model as skewed dispersion may confound findings from PERMANOVA.

Differential abundance analyses

Microbial counts at the genus level were filtered for at least 10% prevalence and centre log-ratio (clr) transformed as recommended by Nearing et al.[19]. Differential abundance analyses were done using the compositional data analysis method, namely with ALDEx2, ANCOM-BC2, MaAsLin2, LinDA and Zicoseq. Bacterial taxa were deemed to be significantly different if the FDR-corrected p-value was <0.05 in more than two algorithms.

Random Forest modelling

Supervised machine learning using random forest models were tested on a total of 883 samples after filtering out samples with missing data or failed filtering QC. The R packages 'caret' (v. 6.0-94) and 'mikropml' (v. 1.6.1) were used to train and test random forest models using an 80:20 training:testing split with 5-fold cross-validation repeated five times (the "xgbTree" method was used with the "repeatedCV" option set to 5 repetitions). The R packages 'mikropml' (v. 1.6.1) and 'MLeval' (v. 0.3) were used to calculate F1, AUROC, precision, sensitivity, specificity and generate plots.

In vitro validation of microbial counts

The qPCR assays were performed with QuantiNova SYBR Green RT-PCR Kit (Qiagen) using the Applied Biosystems Real-Time PCR System. Twenty nanograms per microliter (ng/µL) of DNA extracted from tumour samples were used as the template for amplification of the V3V4 region using the forward primer (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′) and reverse primer (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3′) sequences obtained from Klindworth et al.[20]. Reaction mixtures consisted of 10 µL master mix, 3 µL each of forward and reverse primers, and 4 µL of DNA template. *Escherichia coli* gDNA was employed as a positive control, OKF6/TERT1 gDNA as a negative control, and water was used as a blank control. Cycles consisted of the following regime: 2 min at 50 °C, 10 min at 95 °C, 40 cycles of 15 s at 95 °C and 30 s at 60 °C, followed by 15 s at 95 °C, 1 min at 60 °C, 30 s at 95 °C, and 15 s at 60 °C for melt curve analysis. A total of 5 µL of the final qPCR amplicons were subjected to agarose gel electrophoresis in a 2% gel at 100 volts for 30 min and visualised under ultraviolet (UV) transillumination on the Azure Biosystems Imaging System. *E. coli* gDNA was serially diluted and ran in triplicates on the qPCR system. The measured threshold cycle (Ct) values were plotted against calculated copy numbers for each reaction. Ct values from the V3V4 qPCR analyses were used to estimate copy numbers of total bacteria present in tumour samples based on the standard curve. Estimated copy numbers were then compared with microbial counts of corresponding samples and used to generate a correlation plot.

Ethical approval

Patient recruitment and sample collection for the MyBrCa cohort was reviewed and approved by the Independent Ethics Committee, Ramsay Sime Darby Health Care (Reference no: 201109.4 and 201208.1), as well as the Medical Ethics Committee of the University Malaya Medical Centre (Reference no: 842.9). All research was performed in accordance with relevant guidelines and regulations. Written informed consent to participation in research was given by each individual patient.

Results

Alpha, beta diversity and most prevalent genera in the Malaysian breast tumour microbiome

Using our discovery cohort (*n* = 558), bacteria read counts were converted to relative abundance to observe the overall distribution of each taxonomy when grouped by PAM50 subtype (Fig. 1A). Proteobacteria, Firmicutes, Actinobacteria, and Heunggongvirae were the most dominant phyla of the breast tumour microbiota. There was significant heterogeneity, where some phyla, such as Acidobacteria, were observed in some samples but were completely absent in others. The top ten most abundant genera in our discovery cohort by median read counts

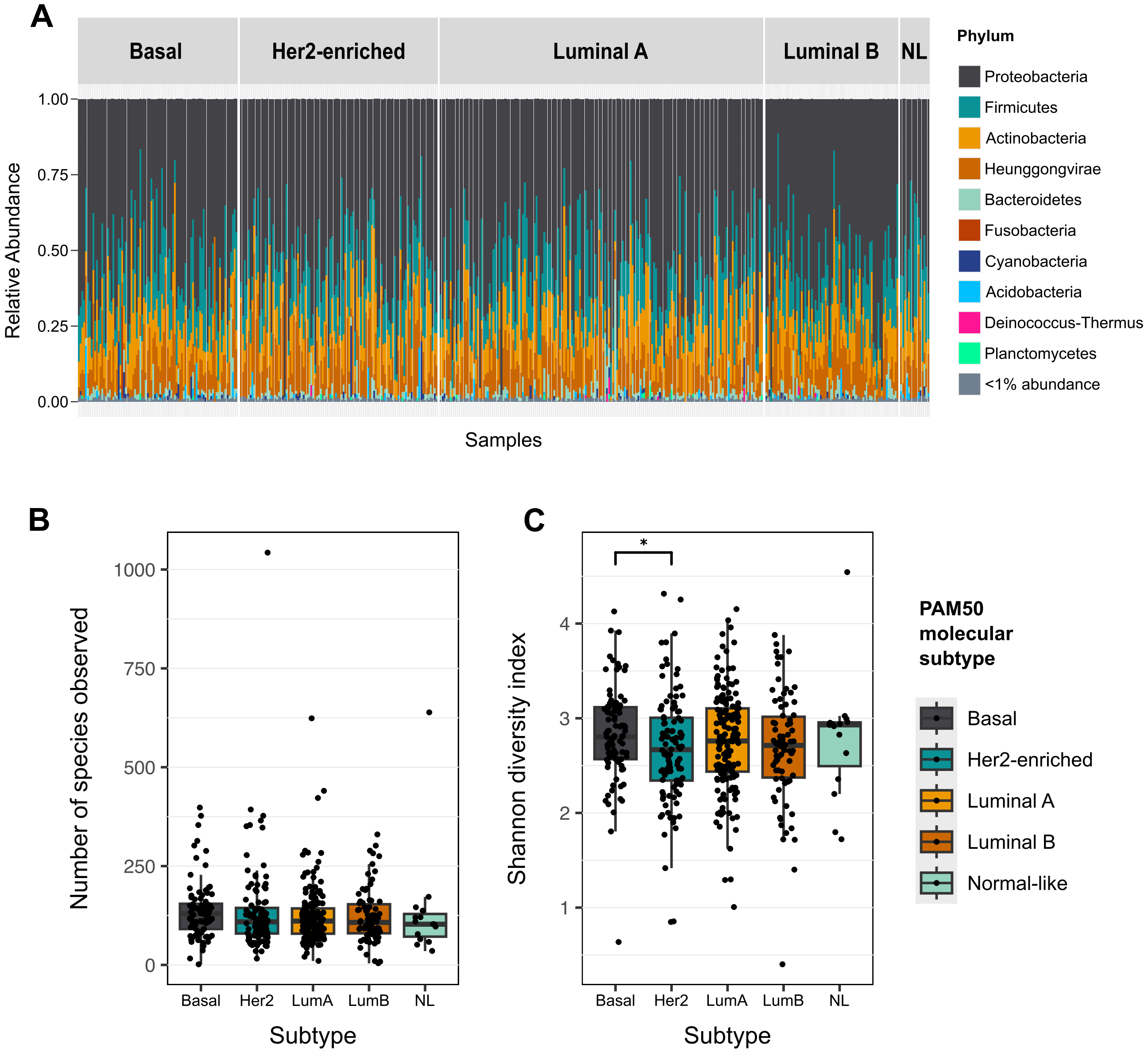

Figure 1.. Detection of intratumoral microbiota from transcriptomic sequencing of Asian breast cancer samples. (**A**) Relative abundance of microbial phyla detected in breast cancer samples from the MyBrCa discovery cohort (n = 558), grouped by PAM50 molecular subtype (NL = normal-like). Also shown are the number of species observed (**B**) and the Shannon diversity index (**C**) for detected intratumoral microbiota across different breast cancer molecular subtypes.

were *Pseudomonas*, *Siphoviridae*, *Bacillus*, *Escherichia*, *Klebsiella*, *Streptomyces*, *Priestia*, *Cutibacterium*, *Serratia*, and *Acinetobacter*.

The intra-group diversity of breast tumour microbiota was mostly homogenous when comparing between molecular subtypes. A slightly higher diversity was observed in Basal subtype when using observed number of species as alpha diversity metric, although it did not reach statistical significance (Fig. 1B, $p > 0.05$). The diversity of the Basal subtype microbiota was significantly higher than the Her2 subtype when compared using the Shannon index (Fig. 1C, $p = 0.027$).

We calculated relative abundance of tumour microbiome and multi-dimensional scaling (MDS) using the Bray–Curtis index to find differences between group (beta-diversity). Unsupervised coordination using PCoA revealed no distinct patterns by PAM50 subtype (Supplementary Fig. 2). We also plotted a Shepherd's stress plot to measure how well the reduced dimensions reflect the original dissimilarity structure. Relative stress, which is a measure of goodness of fit in MDS and preferably lower value, was 0.29 (Supplementary Fig. 3), indicating that the MDS was a decent fit to the original dissimilarity structure.

In order to examine dissimilarity between groups (inter-group diversity) and the variance contributed by each covariate, we conducted a PERMANOVA analysis, which revealed significant differences between individuals with high versus low IFNγ immune scores[21] (*F*-statistic = 2.431, $p = 0.015$, Table 1). Dissimilarity contributed by immune scores remained significant when substituted by other immune scores such as Bindea[22] and ESTIMATE[23]. We also calculated homogeneity using PERMDISP to ensure that differences in group was due to variance and not sample dispersion. All variables examined had homogenous dispersion, with the exception of age at diagnosis (*F*-statistic = 1.715, $p = 0.003$, Supplementary Table 1). It is interesting to note that age at diagnosis explained 19% of variance observed for dispersion. This is expected because patients in this cohort range from 22 – 85 years old, thus resulting in high dispersion.

Inter-group diversity was visualized using supervised ordination with distance-based Redundancy Analysis (dbRDA) which reflected similar findings to PERMANOVA (Figure 2). The figure shows that two axes chosen, RDA1 and RDA2 explained the highest tumour microbiome variance in a multi-dimensional data at 30.9% and 25.2%. The IFNγ immune score had the most significant effect on the variance observed, as confirmed in PERMANOVA.

Differential abundance analyses of immune scores

Given the previous observation that immune scores had the most significant association with the variance observed in microbial abundance, we investigated which bacteria may be associated with differences in immune scores using differential abundance analysis. Immune scores included in the analysis were Bindea, ESTIMATE, IMPRES, CD8, and IFNγ immune scores as scored in Pan et al.[5]. Immune scores were grouped into high and low by their median. Multiple algorithms, namely ALDEx2, ANCOM-BC2, MaAslin2, Zicoseq, and LinDA, were utilized to search for a consistent pattern while avoiding algorithmic bias towards the identification of differentially abundant bacteria taxa[19]. Significant findings were defined as those genera with FDR-adjusted *p*-value < 0.05 by two or more algorithms.

These analyses showed that *Sulfidibacter* was significantly increased in patients across most high immune score groups (Bindea, ESTIMATE, CD8 and IFNγ; *p*-value < 0.05, Table 2). Additionally, *Priestia* and *Pseudoalteromonas* were significantly increased in IFNγ high groups, while *Bacillus* was significantly increased in patients categorized into low IMPRES score group across at least two separate algorithms.

Validation of significant associations in a validation cohort

Samples that were sequenced in a later batch were used as a validation cohort ($n = 419$). In order to validate our previous finding of an association between the microbial abundance of specific bacterial genera with immune scores, we compared the normalized abundance of *Sulfidibacter*, *Priestia*, and *Pseudoalteromonas* between samples with high versus low immune scores.

We found that *Sulfidibacter* was significantly higher in abundance among the higher immune score groups for Bindea, ESTIMATE, and IFNγ (Fig. 3, *t*-test $p < 0.05$), but not CD8 ($p = 0.38$), in our validation cohort. Additionally, *Priestia* was also significantly higher in patients with high IFNγ scores in our validation cohort ($p < 0.0001$). However, contrary to our discovery cohort, *Pseudoalteromonas* was not significantly different

	df	SS	F	Pr(>F)	Total variance	Explained variance
Model	12	0.87	1.03	0.391	27.69	0.031
PAM50 subtype	4	0.31	1.12	0.283	27.69	0.011
Age at diagnosis	1	0.085	1.21	0.227	27.69	0.0031
Ethnicity	3	0.14	0.68	0.891	27.69	0.0052
Stage	1	0.048	0.69	0.752	27.69	0.0017
Grade	1	0.078	1.11	0.323	27.69	0.0028
Chemotherapy	1	0.060	0.86	0.535	27.69	0.0022
IFNγ group	1	0.17	2.43	0.015	27.69	0.0062
Residual	382	26.83	NA	NA	27.69	0.97

Table 1.. PERMANOVA analysis of the tumour microbiome in Malaysian breast cancer patients

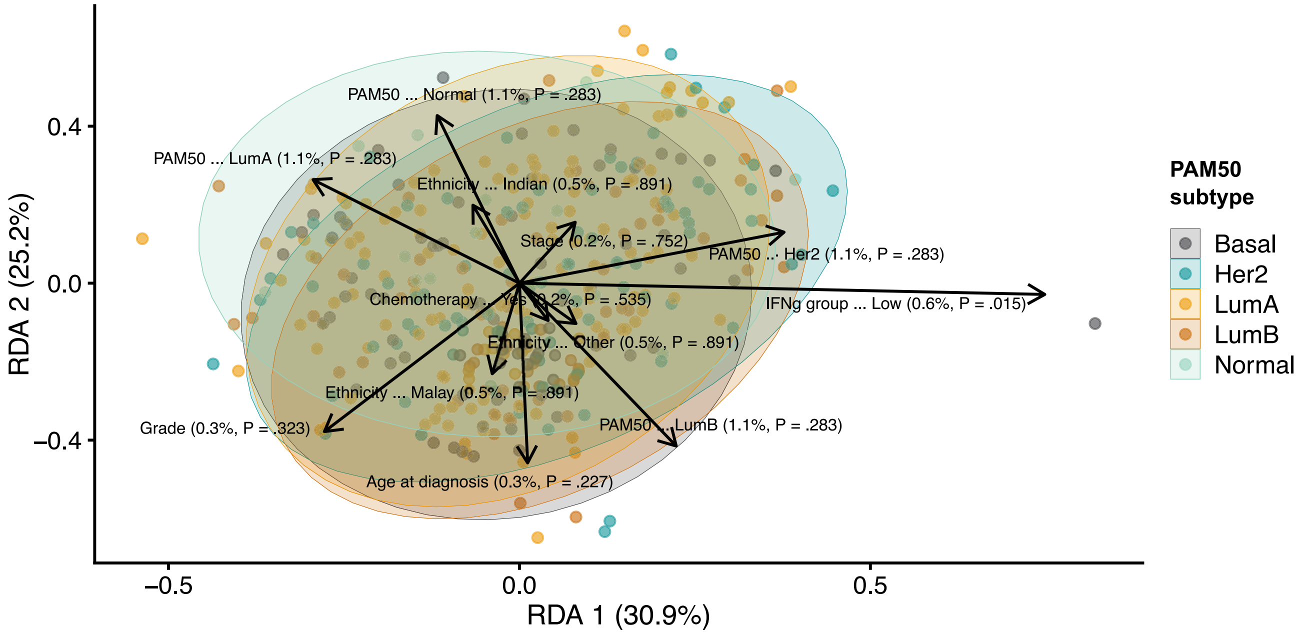

Figure 2.. Inter-group diversity of detected intratumoral microbiota, as visualized using supervised ordination with distance-based Redundancy Analysis (dbRDA) against clinical and molecular variables. Clinical and molecular variables included in the analysis were PAM50 molecular subtype, IFN-γ scores (grouped according to median value), ethnicity, chemotherapy treatment (received or not), age at diagnosis, cancer stage, and tumour grade.

between IFNγ high and low groups ($p = 0.16$). Overall, the results from our validation cohort confirmed most but not all of the associations between bacterial abundance and immune scores from the discovery cohort.

Random Forest prediction of immune scores from microbiome data

We used machine learning to explore the possibility of utilising the tumour microbiota to predict samples with high or low immune scores. A 5-fold cross-validation random forest model was used with an 80-20 split between the training and testing dataset. The random forest model successfully predicted immune high and immune low groups in our full dataset ($n = 883$), with an area under the ROC curve (AUC-ROC) of 0.80 for IFNγ, 0.78 for Bindea, 0.72 for ESTIMATE, 0.72 for CD8, and 0.60 for IMPRES (Table 3, Figure 4A). Across all five immune scores analysed, the random forest models of the tumour microbiota found an association with immune scores that was significantly better than chance (AUC-ROC 95% CI > 0.5) and with moderately high sensitivity and specificity in most cases except for IMPRES. The random forest model with the best predictive performance was for IFNγ scores, with an area under the precision-recall curve (AUC-PR) of 0.76 (Fig. 4B), and the top three features that contributed to the random forest binary classification model for IFNγ were *Sulfidibacter, Prestia,* and *Erythrobacter* (Fig. 4C). Importantly, the tumour microbiome was still significantly associated with IFNγ scores even when *Sulfidibacter* alone or *Sulfidibacter* and *Priestia* were dropped from the training data (AUROC of 0.72 [95% CI 0.70–0.76] and 0.70 [95% CI 0.67–0.73] respectively, Supplementary Table 2), suggesting that this association may be robust.

In vitro validation

We also sought to validate the existence and overall abundance of the tumour microbiome in our samples using orthogonal methods. Thus, we conducted an in vitro validation of microbial abundance using qPCR amplification of the bacterial 16S V3V4 region of 20 randomly-selected samples from our cohort. The estimated total microbial copy numbers derived from qPCR were then compared with microbial read counts derived from RNA sequencing in order to determine their correlation (Fig. 5). Both Spearman's and Pearson's correlation revealed moderately strong associations between the two variables, with correlation coefficients of 0.513 ($p = 0.020$) and 0.7104 ($p = 0.00045$) respectively, suggesting that our overall per-sample microbial read counts derived from RNA sequencing were reliable.

Discussion

In this study, we sought to characterize the tumoral microbiota of Asian breast cancer patients to understand its association with molecular subtypes and immune scores. We used a compositional data analysis method involving five algorithms to analyze microbial read counts derived from 558 RNA-seq samples, followed by validation with a separate cohort of 419 samples as well as qPCR of 20 randomly-selected samples. Our findings suggest a lack of association between the intra-tumoral microbiome and most clinical variables, but also suggest a potential association between the intra-tumoral microbiome and immune scores in our Asian cohort.

Immune score	Taxa	Effect size/log fold change	FDR q-value	Test used
Bindea	*Sulfidibacter*	− 0.31	0.002	ALDEx2
	Sulfidibacter	− 0.70	0.0032	ANCOM-BC2
	Bifidobacterium	− 0.26	0.0092	
	Microbacterium	0.0021	0.031	
	Kocuria	0.095	0.032	
	Pseudolysobacter	0.099	0.032	
	Sulfidibacter	0.028	0.0024	MaAslin2
	Sulfidibacter	–	0.001	Zicoseq
	Sulfidibacter	− 1.32	1.51E−07	LinDA
	Mycobacteroides	0.26	0.036	
	Kinneretia	0.36	0.036	
ESTIMATE	*Sulfidibacter*	− 0.311	2.10E−06	ALDEx2
	Dietzia	0.405	0.00338	ANCOM-BC2
	Stenotrophomonas	− 0.497	0.00338	
	Curtobacterium	− 0.391	0.00647	
	Pseudonocardia	0.334	0.0145	
	Pseudolysobacter	0.372	0.0145	
	Paraburkholderia	0.309	0.0146	
	Sulfidibacter	− 0.428	0.0403	
	Alcanivorax	0.306	0.0403	
	Gordonia	0.352	0.0421	
	Sulfidibacter	− 0.0281	0.00085	MaAslin2
	Sulfidibacter	–	0.00050	Zicoseq
	Sulfidibacter	− 1.157	9.27E−06	LinDA
	Priestia	− 0.816	0.0260	
IMPRES	*Bacillus*	0.0631	0.0033	MaAslin2
	Bacillus		0.001	Zicoseq
CD8	*Sulfidibacter*	− 0.263	0.002	ALDEx2
	Sulfidibacter	–	–	LinDA
IFNγ	*Sulfidibacter*	− 0.441	0	ALDEx2
	Priestia	− 0.222	0.024	
	Sulfidibacter	− 0.760	6.28E−06	ANCOM-BC2
	Curtobacterium	− 0.476	0.00312	
	Dietzia	0.395	0.00696	
	Dolosigranulum	0.342	0.00696	
	Rhizobium	− 0.336	0.00850	
	Micrococcus	− 0.495	0.0111	
	Priestia	− 0.547	0.0152	
	Pseudonocardia	0.288	0.0152	
		Coefficient in low group		
	Sulfidibacter	− 0.039	0	MaAslin2
	Erythrobacter	− 0.007	0.026	
	Pseudoalteromonas	− 0.005	0.026	
	Sulfidibacter	–	–	Zicoseq
	Mycobacterium	–	–	
	Priestia	–	–	
		Log$_2$ fold change		
	Sulfidibacter	− 1.84	3.15E−16	LinDA
	Priestia	− 1.01	0.000831	
	Pseudoalteromonas	− 0.358	0.0193	

Table 2.. Results for differential abundance analysis for each immune score. Also indicated are the the effect size/log fold change and the test used. A negative effect size/log fold change indicates that the taxa was enriched in the group with high immune scores.

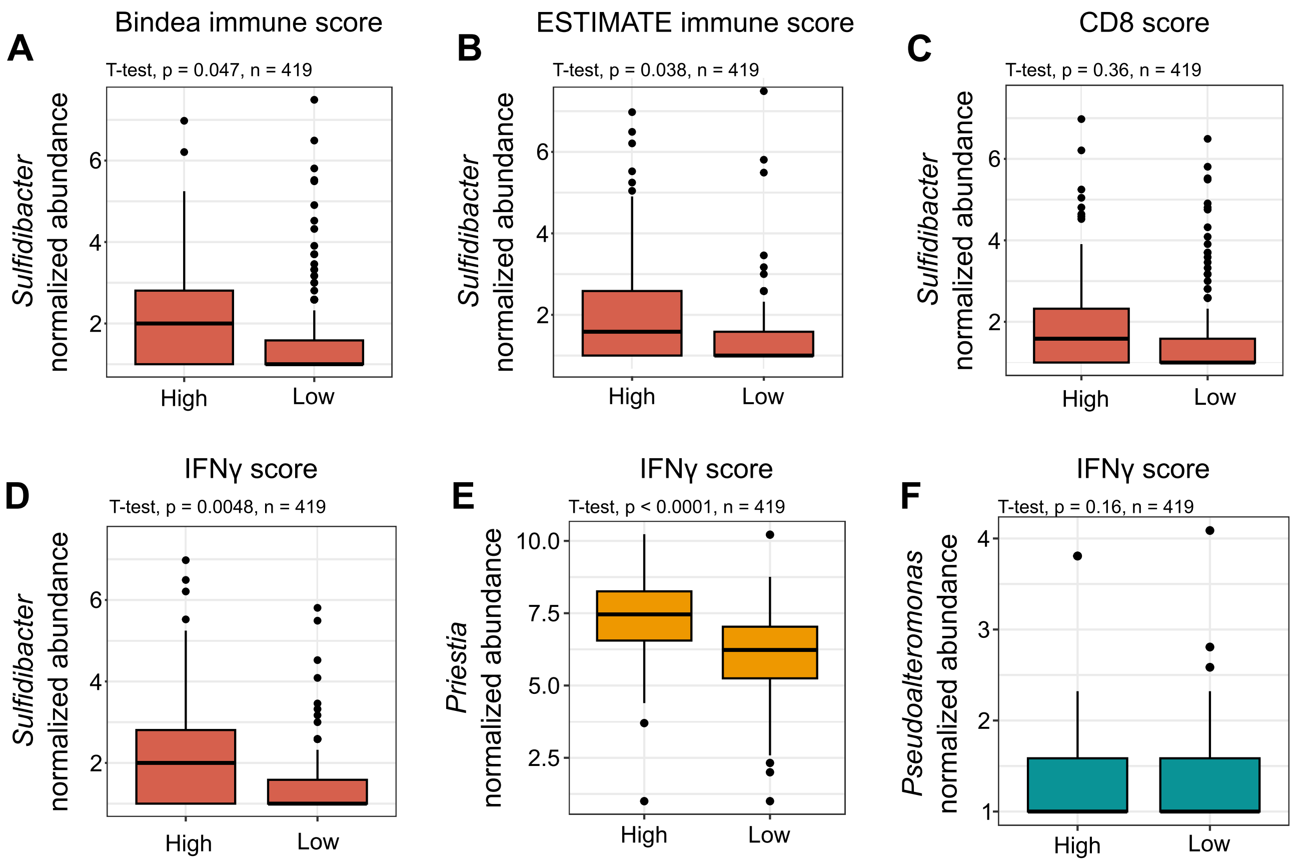

Figure 3.. Association of intratumoral microbiota with immune scores in a validation cohort of Asian breast cancer samples (n = 419). (**A**–**D**) Comparison of RNAseq-derived normalized abundance scores for *Sulfidibacter* between samples with high versus low immune scores according to the median value, for Bindea, ESTIMATE, CD8, and IFN-γ immune scores, respectively. (**E**,**F**) Comparison of RNAseq-derived normalized abundance scores for *Priestia* (**E**) and *Pseudoalteromoas* (**F**) between samples with high versus low IFN-γ immune scores, according to the median value.

Immune score	F1 Score	Area under the ROC curve (95% CI)	Precision (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
IFNγ	0.75	0.80 (0.77–0.83)	0.67 (0.63–0.70)	0.87 (0.83–0.89)	0.61 (0.57–0.66)
Bindea	0.72	0.78 (0.75–0.81)	0.70 (0.66–0.74)	0.74 (0.70–0.78)	0.70 (0.65–0.73)
ESTIMATE	0.68	0.72 (0.69–0.75)	0.65 (0.61–0.69)	0.70 (0.66–0.74)	0.63 (0.59–0.67)
CD8	0.65	0.72 (0.69–0.75)	0.65 (0.61–0.69)	0.64 (0.60–0.69)	0.68 (0.64–0.72)
IMPRES	0.27	0.60 (0.54–0.66)	0.17 (0.14–0.21)	0.65 (0.56–0.73)	0.55 (0.51–0.58)

Table 3.. Random forest prediction metrics for prediction of immune scores using intratumoral microbiome relative abundance scores (n = 883).

We observed a largely homogenous intra-group diversity in the Malaysian breast tumor microbiome across PAM50 subtypes, except for the basal subtype which had a significantly more diverse microbiota composition compared to the HER2-enriched subtype. The homogeneity observed is in line with Desalegn et al.[24] who reported no significant differences in tumour microbiota between PAM50 subtypes among Ethiopian breast cancer patients. Kim et al.[25] reported similar findings in a Korean cohort, additionally showing two distinct clusters independent of subtypes associated with regional recurrence free survival. Other studies have reported distinct microbiome compositions between tumour and normal adjacent tissue samples[12,26,27], which is expected given that comparisons between healthy and diseased microbiomes have consistently reported lower microbial diversity in the latter[28].

However, the significant difference in microbiome diversity between the basal and HER2-enriched subtype has not been reflected in current literature. Interestingly, Chen et al.[29] reported that Asian breast cancer patients are less likely to have luminal A and basal subtypes but more likely to have luminal B and HER2-enriched subtypes than Western patients. Tumour microbiomes tend to be less well-characterized compared to the gut microbiome. This gap in knowledge is further exacerbated by the lack of Asian-centric cohorts[25,30]. Furthermore,

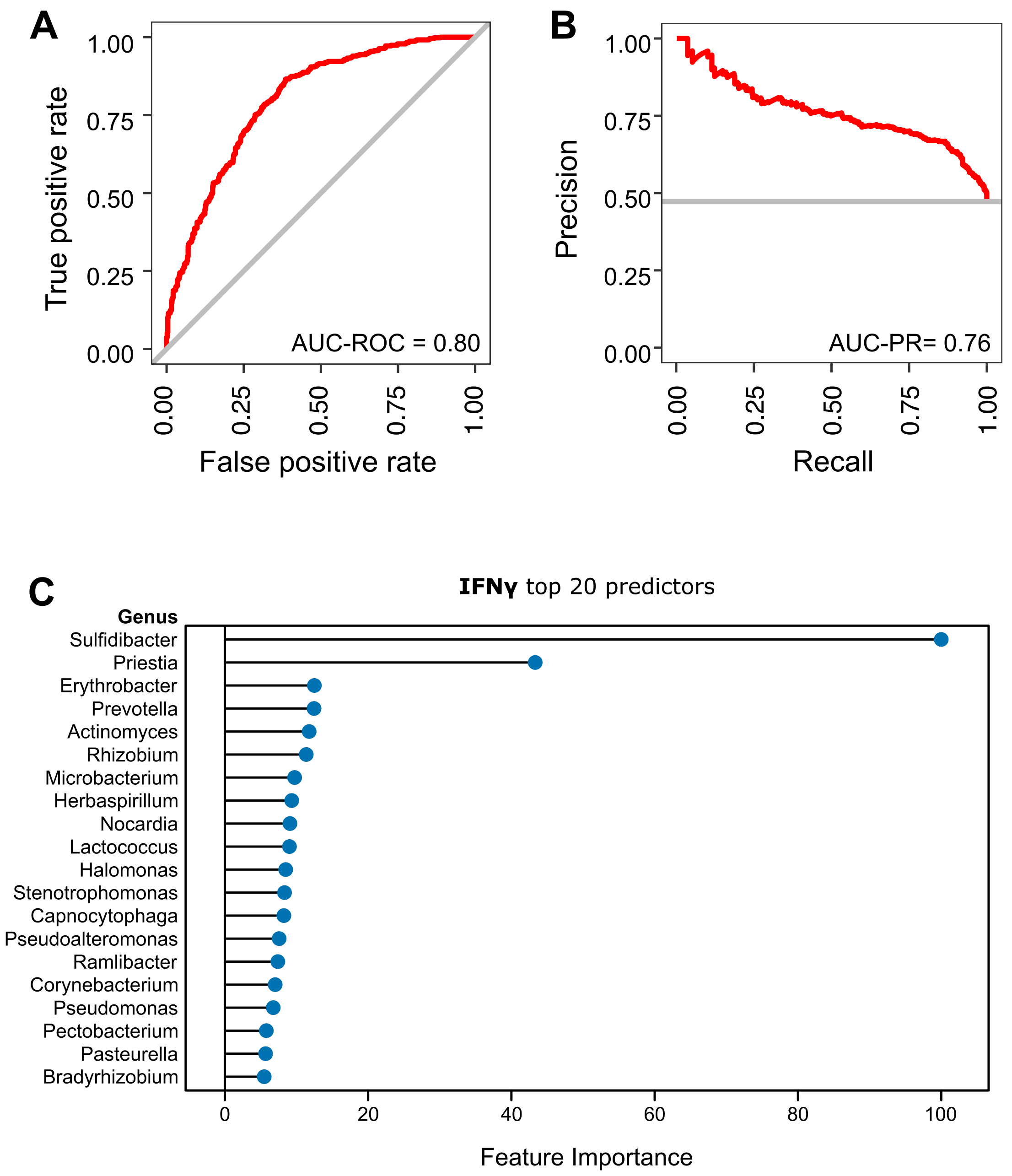

Figure 4.. Random forest prediction of IFN-γ scores using intratumoral microbial abundance. Receiver operating characteristics (**A**) and precision-recall curve (**B**) for a random forest model trained to predict IFN-γ scores using intratumoral microbial abundance. (**C**) Top 20 most important features (abundance of specific microbial genera) used by the random forest model to predict IFN-γ scores.

tumour microbiome studies with multiethnic cohorts tend to have a relatively low representation of Asians[12,13]. Considering the sample size of our cohort, it is possible that the observed differences in microbiome diversity between basal and HER-2 enriched subtypes could be specific to Asian populations but this requires further validation.

The results of our inter-group diversity analysis further revealed that variation in the Malaysian breast tumour microbiome was significantly associated with immune scores, while molecular subtype, age at diagnosis, cancer stage, tumour grade, ethnicity, and treatment type had no association with the variation found in the Malaysian breast tumour microbiome.

There is evidence that microbes can interact with cells patrolling the tumour microenvironment, notably close interactions with immune cells possibly affecting tumour inhibition and proliferation[31]. Microbes have been found intracellularly in both cancer and immune cells[9]. In other cancers such as pancreatic ductal adenocarcinoma, a uniquely diverse composition of tumour microbiome distinct from that of adjacent

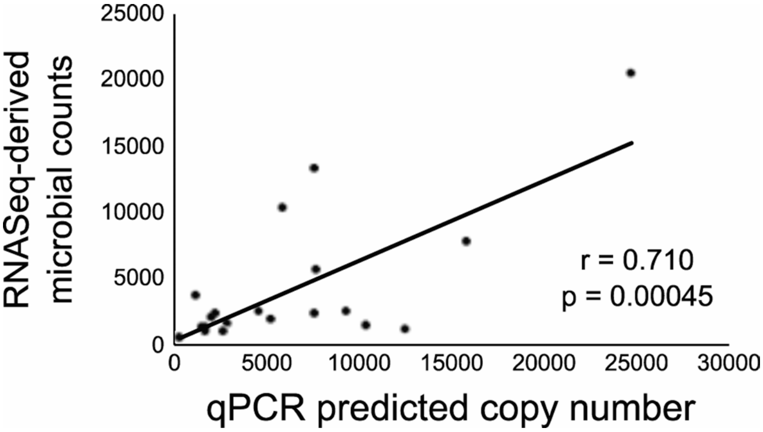

Figure 5.. Predicted copy number for the combined intratumoral microbiome, as determined by qRT-PCR amplification of bacterial 16S V3V4, compared to total transcript counts for all detected microbiota derived from RNAseq, for 20 Asian breast cancer tumour samples from the MyBrCa cohort.

healthy pancreatic tissue was found to be associated with more sustained CD8$^+$ T cell response in the tumor microenvironment[8].

Increased immune cell response and variations in immune scores have also been attributed to microbe-derived metabolites present in the tumour immune microenvironment (TIME)[31]. Short chain fatty acids (SCFAs), such as butyric acid, are known microbe-derived metabolites which can accumulate within tumors and inhibit histone deacetylases (HDACs), referring to chromatin regulatory factors expressed abnormally in a variety of human cancers[32]. Butyrate-mediated HDAC inhibition causes the upregulation of transcriptional regulator ID2, triggering the IL-12R signaling pathways in CD8$^+$ T cells[33]. This results in an increased CD8$^+$ T cell density and activation in the TIME.

In the case of IFNγ immune score, studies have reported that some bacterial genera can promote IFNγ secretion, including a recently defined community of 11 bacteria that induced IFNγ production preferentially in CD8$^+$ T cells in the absence of immunotherapy[34,35]. IFNγ secretion in the TIME have also been linked to *Bifidobacterium*[36]. One such metabolite is inosine, a purine metabolite which induces naïve T cells to differentiate into CD4$^+$ Th1, leading to increased CD8$^+$ T-cell infiltration and IFNγ secretion, especially in combination with PD-L1 blockade[32,37]. It is worth noting that while IFNγ is classically associated with anti tumour effects, IFNγ can upregulate proliferative signals and allow tumour cells to escape recognition by immune cells under certain conditions[38].

Our differential abundance analyses using center log-ratio transformed counts and five different algorithms showed that *Sulfidibacter* was significantly increased in patients with higher immune scores, including Bindea, ESTIMATE, and IFNγ. Additionally, *Priestia* was also more abundant in patients with high IFNγ scores in both our discovery and validation cohorts.

The presence of *Sulfidibacter* in the tumour microbiome was unexpected as it is a novel marine bacterium first isolated and identified from corals[39]. To date, Wang et al.[39] is the only publication available which characterizes *Sulfidibacter*. However, given it was proposed as a species of Acidobacteria and members of this phylum are typically associated with aquatic, terrestrial, and extreme environments, *Sulfidibacter* is not expected nor likely to appear in human species. Hence, it is possible that its presence in our data is the result of taxonomic misclassification due to database contamination by human reads or other contamination instead of a true biological signal[40].

Priestia is another marine bacterium previously reported as an arginase producer, an enzyme with potential in cancer treatment by arginine deprivation therapy[41]. *Priestia* was previously identified in a Slovakian breast tumour cohort by Hadzega et al.[42], who conducted transcriptomic sequencing to investigate the breast tumour microbiome and found that *Priestia* was enriched in breast tumours from patients compared to normal tissues from cancer-free women.

Moving forward, our results may have implications for future treatment strategies to modulate IFNγ in the TIME via manipulation of the tumour microbiome. Already, engineered bacteria injected at tumour sites have been found to trigger IFNγ expression through a cascade of pathways that increases anti-tumour effects[43]. Similarly, Kim et al.[44] demonstrated the use of gram-negative bacteria outer membrane vesicles to induce anti-tumour effects through the production of anti-tumour cytokines such as IFNγ and CXCL10.

One of the strengths of our study, aside from the sizeable cohort, is the analysis strategy used to mediate batch effect. Batch effect has been and will continue to be a major issue with the rise of big data and large microbiome cohorts. Several strategies to correct it have been reported in literature including conditional quantile regression[45], MBECS[46], Limma[47], and ComBat[48]. Still, there persists the question of whether these strategies could overcorrect data to the point of distorting data dispersion, resulting in the detection of false positive signals or the masking of true positive signals[49]. To avoid such data distortion, we adapted a strategy from Sepich-Poore et al.[49] where we used one sequencing batch as an exploratory cohort and another batch as an independent validation cohort.

Our study does have some limitations. The dataset was not initially designed for microbiome investigations, and thus, there is a lack of microbiome controls to rule out environmental contamination. We attempted to reduce the effect of this on our findings by applying appropriate prevalence filtering, testing differential association on five different algorithms, and incorporating the use of a sizable validation cohort. We have also conducted orthogonal validation via qPCR on tumour DNA. However, we cannot completely rule out the presence of contaminants or false positives in our data. Additionally, the inclusion of other omics such as metabolomics, genomics, and gut metagenomics could provide more insights into understanding of the human microbiome and its role in association with cancer.

Data availability

Whole exome sequencing and RNA-seq data used in this study are accessible from the European Genome-phenome Archive under accession numbers EGAS00001006518 (https://ega-archive.org/studies/EGAS00001006518) and EGAS00001004518 (https://ega-archive.org/studies/EGAS00001004518). Access to controlled patient data will require the approval of the Data Access Committee. Further information is available from the corresponding author upon request.

Code availability

Code used to produce the analysis are publicly available on GitLab at https://gitlab.com/li-fangyeo/mybrca-tumourmicrobiome/-/tree/76b737f614f9c976eabd9cde62160e8682238343/.

Received: 14 April 2025; Accepted: 11 August 2025
Published online: 24 August 2025

References

1. Breast Cancer Association Consortium. Breast cancer risk genes—association analysis in more than 113,000 women. *N. Engl. J. Med.* **384**, 428–439 (2021).
2. Mertens, E. et al. Understanding the contribution of lifestyle in breast cancer risk prediction: a systematic review of models applicable to Europe. *BMC Cancer* **23**, 687 (2023).
3. Mao, X. et al. *BMC Cancer* **23**, 644 (2023).
4. Martini, R. et al. African ancestry-associated gene expression profiles in triple-negative breast cancer underlie altered tumor biology and clinical outcome in women of african descent. *Cancer Discov.* **12**, 2530–2551 (2022).
5. Pan, J.-W. et al. The molecular landscape of Asian breast cancers reveals clinically relevant population-specific differences. *Nat. Comm.* **11**, 6433 (2020).
6. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
7. Routy, B. et al. Gut microbiome influences efficacy of PD-1 based immunotherapy against epithelial tumors. *Science* **359**, 91–97 (2018).
8. Riquelme, E. et al. Tumor microbiome diversity and composition influence pancreatic cancer outcomes. *Cell* **178**, 795–806 (2019).
9. Nejman, D. et al. The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science* **368**, 973–980 (2020).
10. Galeano Niño, J. L. et al. Exploring breast tissue microbial composition and the association with breast cancer risk factors. *Breast Cancer Res.* **25**, 82 (2023).
11. Fu, A. et al. Tumor-resident intracellular microbiota promotes metastatic colonization in breast cancer. *Cell* **185**, 1356–1372 (2022).
12. German, R. et al. Exploring breast tissue microbial composition and the association with breast cancer risk factors. *Breast Cancer Res.* **25**, 82 (2023).
13. Parida, S. et al. Concomitant analyses of intratumoral microbiota and genomic features reveal distinct racial differences in breast cancer. *NPJ Breast Cancer* **9**, 4 (2023).
14. Pan, J.-W. et al. Clustering of HR+/HER2− breast cancer in an Asian cohort is driven by immune phenotypes. *Breast Cancer Res.* **26**, 67 (2024).
15. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
16. Wood, D. E. et al. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
17. Lu, J. et al. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
18. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
19. Nearing, J. T. et al. Microbiome differential abundance methods product different results across 38 datasets. *Nat. Comm.* **13**, 342 (2022).
20. Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
21. Ayers, M. et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Investig.* **127**, 2930–2940 (2017).
22. Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
23. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Comm.* **4**, 2612 (2013).
24. Desalegn, Z. et al. Human breast tissue microbiota reveals unique microbial signatures that correlate with prognostic features in adult ethiopian women with breast cancer. *Cancers (Basel).* **15**, 4893 (2023).
25. Kim, H. E. et al. Microbiota of breast tissue and its potential association with regional recurrence of breast cancer in Korean women. *J. Microbiol. Biotechnol.* **31**, 1643–1655 (2021).
26. Tzeng, A. et al. Human breast microbiome correlates with prognostic features and immunological signatures in breast cancer. *Genome Med.* **13**, 60 (2021).
27. Banerjee, S. et al. Prognostic correlations with the microbiome of breast cancer subtypes. *Cell Death Dis.* **12**, 831 (2021).
28. Kriss, M. et al. Low diversity gut microbiota dysbiosis: drivers, functional implications and recovery. *Curr. Opin. Microbiol.* **44**, 34–40 (2018).
29. Chen, C. H. et al. Disparity in tumor immune microenvironment of breast cancer and prognostic impact: Asian Versus Western Populations. *Oncologist* **25**, e16–e23 (2020).
30. Luo, L. et al. Species-level characterization of the microbiome in breast tissues with different malignancy and hormone-receptor statuses using nanopore sequencing. *J. Pers. Med.* **13**, 174 (2023).

31. Ma, J. et al. The role of the tumor microbe microenvironment in the tumor immune microenvironment: bystander, activator, or inhibitor?. *J. Exp. Clin. Cancer Res.* **40**, 327 (2021).
32. He, Y. et al. Gut microbial metabolites facilitate anticancer therapy efficacy by modulating cytotoxic CD8+ T cell immunity. *Cell Metab.* **33**, 988–1000 (2021).
33. Luu, M. et al. Regulation of the effector function of CD8+ T cells by gut microbiota-derived metabolite butyrate. *Sci. Rep.* **8**, 14430 (2018).
34. Tanoue, T. et al. A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600–605 (2019).
35. Wang, H. et al. Breast tissue, oral and urinary microbiomes in breast cancer. *Oncotarget* **8**, 88122–88138 (2017).
36. Rezasoltani, S. et al. Modulatory effects of gut microbiome in cancer immunotherapy: A novel paradigm for blockade of immune checkpoint inhibitors. *Cancer Med.* **10**, 1141–1154 (2020).
37. Mager, L. F. et al. Microbiome-derived inosine modulates response to checkpoint inhibitor immunotherapy. *Science* **369**, 1481–1489 (2020).
38. Zaidi, M. R. & Merlino, G. The two faces of interferon-γ in cancer. *Clin. Cancer Res.* **17**, 6118–6124 (2011).
39. Wang, G. et al. Comparative genomics reveal the animal-associated features of the acanthopleuribacteraceae bacteria, and description of *Sulfidibacter corallicola* gen. nov., sp., nov. *Front. Microbiol.* **13**, 778535 (2022).
40. Gihawi, A. et al. Major data analysis errors invalidate cancer microbiome findings. *mBio* **14** (2023).
41. Jiao, Y. L. et al. Arginase from *Priestia megaterium* and the effects of CMCS conjugation on its enzymological properties. *Curr. Microbiol.* **80**, 292 (2023).
42. Hadzega, D. et al. Uncovering microbial composition in human breast cancer primary tumour tissue using transcriptomic RNA-seq. *Int. J. Mol. Sci.* **22**, 9058 (2021).
43. Chen, Y. et al. Spatiotemporal control of engineered bacteria to express interferon-γ by focused ultrasound for tumor immunotherapy. *Nat. Comm.* **13**, 4468 (2022).
44. Kim, O. Y. et al. Bacterial outer membrane vesicles suppress tumor by interferon-γ-mediated antitumor response. *Nat. Comm.* **8**, 626 (2017).
45. Ling, W. et al. Batch effects removal for microbiome data via conditional quantile regression. *Nat. Comm.* **13**, 5418 (2022).
46. Olbrich, M. et al. MBECS: microbiome batch effects correction suite. *BMC Bioinform.* **24**, 182 (2023).
47. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
48. Johnson, W. E. et al. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
49. Sepich-Poore, G. D. et al. Robustness of cancer microbiome signals over a broad range of methodological variation. *Oncogene* **43**, 1127–1148 (2024).

Acknowledgements

The authors would like to thank and acknowledge the contributions of the Core Laboratory personnel at Cancer Research Malaysia, the Caldas Lab at the University of Cambridge, as well as the Malaysian Breast Cancer study team in the formation of the dataset used in this study.

Author contributions

L.F.Y. contributed to study design, data collection and processing, and drafting of the manuscript and figures. A.W.Y.L. and J.S.F.C. contributed to the drafting of the manuscript and literature review. P.Y.E.T. contributed to data collection, data analysis, and drafting of the manuscript and figures. B.K.B.L. contributed to study conceptualization and design as well as data collection. J.L. contributed to study design and project supervision. S.H.T. also contributed to study design and project funding, as well as project direction and supervision. J.W.P. contributed to study design, project direction and supervision, data analysis and drafting of the manuscript and figures. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

Funding

Cancer Research Malaysia receives charitable funding from the Khind Starfish Foundation, the Ong Hin Tiang & Ong Sek Pek Foundation, Yayasan PETRONAS, and Yayasan Sime Darby which contributed to the funding of this study. LFY is currently a SYS-LIFE postdoc funded by the European Union´s Horizon Europe Framework programme for research and innovation 2021-2027 under the Marie Skłodowska-Curie grant agreement No 101126611.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15877-x.

Correspondence and requests for materials should be addressed to J.-W.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.