מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

Doctor of Philosophy

Submitted to the Scientific Council of the

Weizmann Institute of Science

Rehovot, Israel

עבודת גמר (תזה) לתואר

דוקטור לפילוסופיה

מוגשת למועצה המדעית של

מכון ויצמן למדע

רחובות, ישראל

By

Lian Narunsky Haziza

מאת

ליאן נרונסקי חזיזה

אפיון המיקוביום בגידולים סרטניים והאפקטים הקליניים שלו

Pan-cancer characterization of the tumor mycobiome and its clinical effects

Advisor:

Prof. Ravid Straussman

Prof. Yitzhak Pilpel

מנחה:

פרופ' רביד שטראוסמן

פרופ' יצחק פלפל

January 2021

שבט תשפ"ב

Acknowledgments

This was an incredible journey that would not have been possible without the support and input from the many people involved in it, both the scientific and non-scientific parts. I would like to start by thanking my two advisors, Prof. Ravid Straussman and Prof. Yitzhak Pilpel. Thank you for the guidance and mentoring, I have learnt a lot from you both, you have shaped me as a scientist. I would like to thank Dr. Ilana Livyatan who was not only my partner in this project but also a friend, for her support and advice throughout. I would like to thank all current and past members of the Straussman lab and Pilpel lab. Specifically I would like to thank Dr. Nancy Gavert, Dr. Deborah Nejman and Omer Asraf for their work on this project, without them this would not have looked the same. I would also like to thank Dr. Orna Dahan for the scientific advice and emotional support whenever it was needed. I would like to thank Dr. Oded Sandler, Dr. Leore Geller, Dr. Tom Kaufman, Aviva Rotter Maskowitz, Dr. Yaara Zwang and Dr. Sivan Kaminski for always listening without you this would have not been possible. I would like to thank my collaborators for the tremendous efforts in making this project what it is today, Dr. Greg Sepich-Poore, Prof. Rob Knight, Guy Amit and Dr. Amir Bashan. I would like to thank my previous advisors Prof. Michal Neeman in my MSc and Prof. Sagiv Shifman in my BSc who encouraged me to take my own path. I would like to thank Dr. Roni Oren and Dr. Liat Alyagor for always knowing what to say to encourage me. Finally, I would like to thank my family. My parents for making me who I am today, for always believing in me, for always encouraging me, for doing all they could to support me. My brothers who were always there to listen. And of course, my partner Roy for believing in me, encouraging me and giving me the time I needed to finish this project. My amazing children Guy, Ronnie and Danna who were all born during this period.

Declaration

This project was done in collaboration with Dr. Ilana Livyatan and Dr. Gregory D. Sepich-Poore. All work on the TCGA data and machine learning on our data were performed by Dr. Gregory D. Sepich-Poore and the Knight lab. Classification of sequencing reads to fungal species in the tumor mycobiome samples was performed by Omer Asraf. Staining of tumor microarrays were performed with Dr. Deborah Nejman, Dr. Nancy Gavert and Ruthie Ariel. Network analysis was performed by Guy Amit and Dr. Amir Bashan.

List of Abbreviations

ITS2- internal transcribed spacer 2; NAT- normal adjacent tissue; PDAC- pancreatic ductal adenocarcinoma; GBM- glioblastoma; TCGA- the cancer genome atlas; WIS- Weizmann; UCSD- University of California San-Diego; IBD- inflammatory bowel disease; ESCC- esophageal squamous cell carcinoma; FISH- fluorescent in-situ hybridization; GMS- Gomori methenamine silver; WGS- whole genome sequencing; HMP- Human Microbiome Project; OS- overall survival; PFS- progression free survival; ML- machine learning; ASV- amplicon sequence variant; TME- tumor microenvironment.

Table of Contents

Abstract

In recent years, cancer-bacteriome interactions are gaining focus in research. However, cancer-associated fungi have rarely been examined. Histological staining for fungi of tissue microarrays revealed the presence of intratumoral fungi, frequently with spatial association with cancer cells and macrophages. This led us to comprehensively characterize the cancer mycobiome (fungal component of the microbiome) within ~1200 human tumors, normal adjacent tissues and normal tissues from a wide range of cancer types, from breast the most common cancer in women to rare cancer types such as glioblastoma. We focused on eight solid tumor types: breast, lung, melanoma, ovary, colon, glioblastoma, pancreas and bone tissues. In addition, we analyzed fungal presence in three additional cohorts including the TCGA dataset of 15,512 sequenced tissue, and blood samples. Together we examined 17,401 samples across 35 cancer types. Fungi are ubiquitous though lowly abundant across all major human cancers, with cancer type-specific compositions. The cancer types differed in their fungal richness as well as fungal load and specific fungal taxa present in the tumors. Fungal profiles were more similar between tumor and normal adjacent tissue of the same tissue source as compared to between tumor types. It is yet to be established whether the fungi detected in the tumors originated in the normal tissue prior to tumor formation or whether the fungi in the normal adjacent tissue originate from the tumor itself. Clinically-focused assessments suggested prognostic and diagnostic capacities of the tissue and plasma mycobiomes. For example, *Malassezia globosa* and Phaeosphaeriaceae were significantly correlated with worse prognosis in either breast or ovarian cancer respectively. We were also able to use the blood fungal compositions to predict accurately cancer patients vs. healthy individuals, even at stage one cancers, indicating a large, still untapped, diagnostic potential of the mycobiome. In addition, comparing intratumoral fungal communities with matched bacteriomes revealed co-occurring, bi-domain ecologies, often with permissive, rather than competitive microenvironments.

Introduction

In recent years the interactions between the human body and the abundance of microorganisms inhabiting it has become a prominent field of research. The human body is a host to many types of microorganisms including bacteria (main component), fungi, viruses and archaea. These organisms can be naïve "bystanders" or play key roles in human health. Research in the field of the bacterial microbiome has undergone a huge leap in recent years (*1–7*), demonstrating that commensal bacteria have many effects on human health and disease. In addition, recent reports identified metabolically-active, immunoreactive, intracellular, cancer type-specific communities of bacteria and viruses living within tumor tissues (*8–20*). Many of these bacteria can render cancer therapies nonfunctional or efficacious (*8–10*, *14*, *15*, *17*, *19*).

In contrast, research into the other microorganisms inhabiting the human body is lagging behind (*21*). Amongst these are the fungi, estimated to be less abundant in the human body than bacteria, with a 1:1000 ratio in the human gut (*22*, *23*). Fungi represent understudied, but critical commensals and opportunistic pathogens that shape host immunity, commonly infecting immunocompromised populations including cancer patients (*24–28*). Recent papers describe the fungal population, or "mycobiome", in different locations of the healthy human body (*29*), including gut (*30–33*), skin (*34*, *35*), oral cavity (*36–38*), and lungs (*39*, *40*). Most of this research is descriptive, characterizing the fungal genera inhabiting these sites in the healthy individual.

In addition, there is accumulating evidence for the role of fungi in disease states. Dysbiosis of the mycobiome was characterized in the gut mycobiome after antibiotic use (*41*), in inflammatory bowel disease (IBD) (*24*), asthma (*41*), in neurodevelopmental disorders (autism and Rett syndrome) (*21*), and in cancer (*42*, *43*). In addition, dysbiosis occurs in HIV patients in the oral cavity, and in cystic fibrosis patients in the lung (*41*). The effects of fungal dysbiosis are not always clear. However, it has been associated with disease severity in IBD patients (*24*). In addition, commensal fungi can modulate the immune system (*44*) demonstrating possible routes by which fungi may affect the diseased states.

To date there are only a handful of papers exploring the tumor mycobiome in several tumor types including colorectal (*42*), breast (*45*, *46*), ovarian (*47*), prostate (*48*) and

pancreatic (*49*) tumors. In all of these papers, fungi were detected in the tumor tissue. In addition, in many cases differences were observed between normal adjacent tissue and tumor tissue, and even between different tumor stages. Most of these studies were done on human samples using a chip array that had probes for only a subset of fungi with pathogenic activity (*45–48*). Next generation sequencing focused on fungal sequencing could expand the characterization of fungi within these tumors.

Two papers demonstrate that fungi can promote tumor growth and progression (*49*, *50*). Zhu et al. (2017) showed, in a mouse model, that chronic fungal infection in the esophagus could promote esophageal squamous cell carcinoma (ESCC). Furthermore, alleviating the fungal infection by antifungal treatment prevented ESCC development (*50*). Aykut et al (2019) demonstrated, in a mouse model, that fungi move from the gut to the pancreas and that the *Malassezia* genera of fungus specifically promotes pancreatic ductal adenocarcinoma (PDAC). The depletion of fungi from the mouse prevented tumor growth (*49*). In addition, fungal load was significantly higher in PDAC in both mice and humans compared to the normal pancreatic tissue. Finally, *Malassezia* was also shown to be one of the dominant species in human PDAC. The mechanism by which fungi promote pancreatic tumor development is through the complement cascade activation (*49*). In addition, Ramirez-Garcia et al. (2016) discussed the mechanisms by which *Candida albicans* may promote tumor development and progression (*51*).

In contrast, fungi may also have anti-tumor activity. Glucans are a major component of the fungal cell wall. Several glucans extracted from fungi were shown to have anti-tumor activity by immune cell activation in mice (*52*, *53*). Hence, fungi in tumors may both promote or inhibit tumor progression.

Whether fungi can act the same way as bacteria and affect different characteristics and stages in tumor progression is mainly unknown, motivating broad characterization of the existence and diversity of the cancer mycobiome. Symbiotic and antagonistic relationships between fungi and bacteria (*54–56*) also motivate studying their interactions in tumors. Herein, we present a comprehensive characterization of the cancer mycobiome in tissues and blood, explore fungal utility in clinically important prognostic and diagnostic cases, and compare fungal communities to matched bacteriomes.

Results

Fungi are detected by multiple staining methods in human tumors

To explore the presence of fungi in human tumors we stained tissue microarrays of melanoma, pancreas, breast, lung, and ovarian cancers (Figures 1, 2). Since no single staining method can detect all fungi in tissues, we used four staining methods: (1) a fungal cell wall-specific anti-β-glucan antibody, whose main caveat is a high false negative rate (the antibody was repurposed from ELISA to IF) (Figure 3), (2) an anti-Aspergillus antibody that also binds several additional fungal species (Figure 3), (3) fluorescence in-situ hybridization (FISH) against three conserved fungal 28S rRNA sequences (*57*) but with selective sensitivity for yeast over hyphal morphologies due to lower hyphae probe penetration (Figures 3C, 4A), and (4) fungal cell wall-specific Gomori methenamine silver (GMS) stain but with high false positive background staining in tissues. Numerous negative controls helped exclude the possibility of false positives (Figures 1, 2). Overall, Anti-Aspergillus was the most common stain detected in tumors with breast, ovary and PDAC tumors having the highest levels of positive tumors, 33%, 25.3% and 15.7%, respectively (Figures 1, 2, 3A). In addition, 21.8% of melanoma tumors and 18.7% of pancreas tumors showed positive β-glucan staining (Figures 1, 2, 3A), while less than 1% of breast, lung, and ovarian tumors were positive for β-glucan staining (Figure 3A). FISH was detected in 12% of PDAC tumors, but was much less abundant in other cancer types (Figures 1A-B, 2D, 3A). GMS staining was difficult to interpret due to high background staining except for rare cases where canonical fungal cells were identified (Figure 4B).

Interestingly, we found a different localization pattern for fungal staining between PDAC and melanoma tumors. In pancreatic tumors, fungal staining (anti- β-glucan & FISH) was mainly evident within cancer cells, whereas melanoma tumors showed macrophage-localized fungal staining (anti- β-glucan and anti-Aspergillus) (Figure 1). In rare cases where canonical fungal cells were identified, they were extracellular (Figure 4).

Overall, with these staining methods we visualized fungi in human tumors; detecting fungal RNA (FISH), cell wall polysaccharides (β-glucan) and proteins (anti-Aspergillus). Yet better methods are needed to overcome the current high false-negative detection rates and to better understand their spatial distribution. We next turned to

exploring the presence and identity of fungal DNA and RNA by qPCR and sequencing methods.

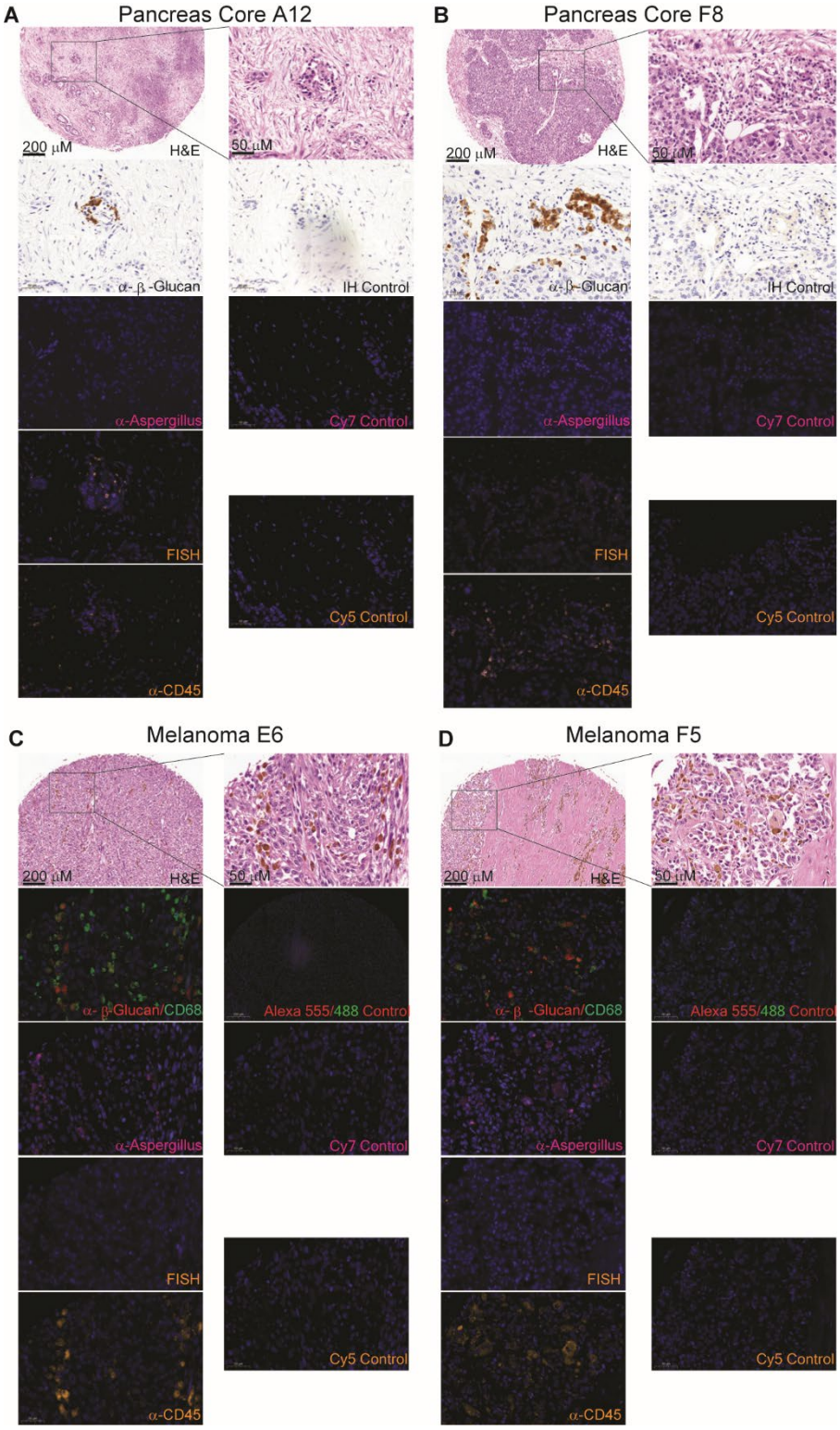


Figure 1. Visualization of fungi in human tumors. (**A-D**) Consecutive slides from a tumor microarray of human pancreatic adenocarcinoma (**A, B**) or Melanoma (**C, D**) were stained with hematoxylin and eosin (H&E), or antibodies against β-glucan, *Aspergillus*, CD45, CD68 or by fluorescence in situ hybridization (FISH) probes against fungal 28*S* rRNA (see Methods). Slides were also stained with only secondary antibodies as a negative control. Two representative cores of each tumor type are presented: PDAC- A12 and F8 (**A, B**), melanoma- E6 and F5 (**C, D**). Scale bar for low magnification: 200μm. Square demarcates the area presented at higher magnification. Scale bar for higher magnification: 50μm.

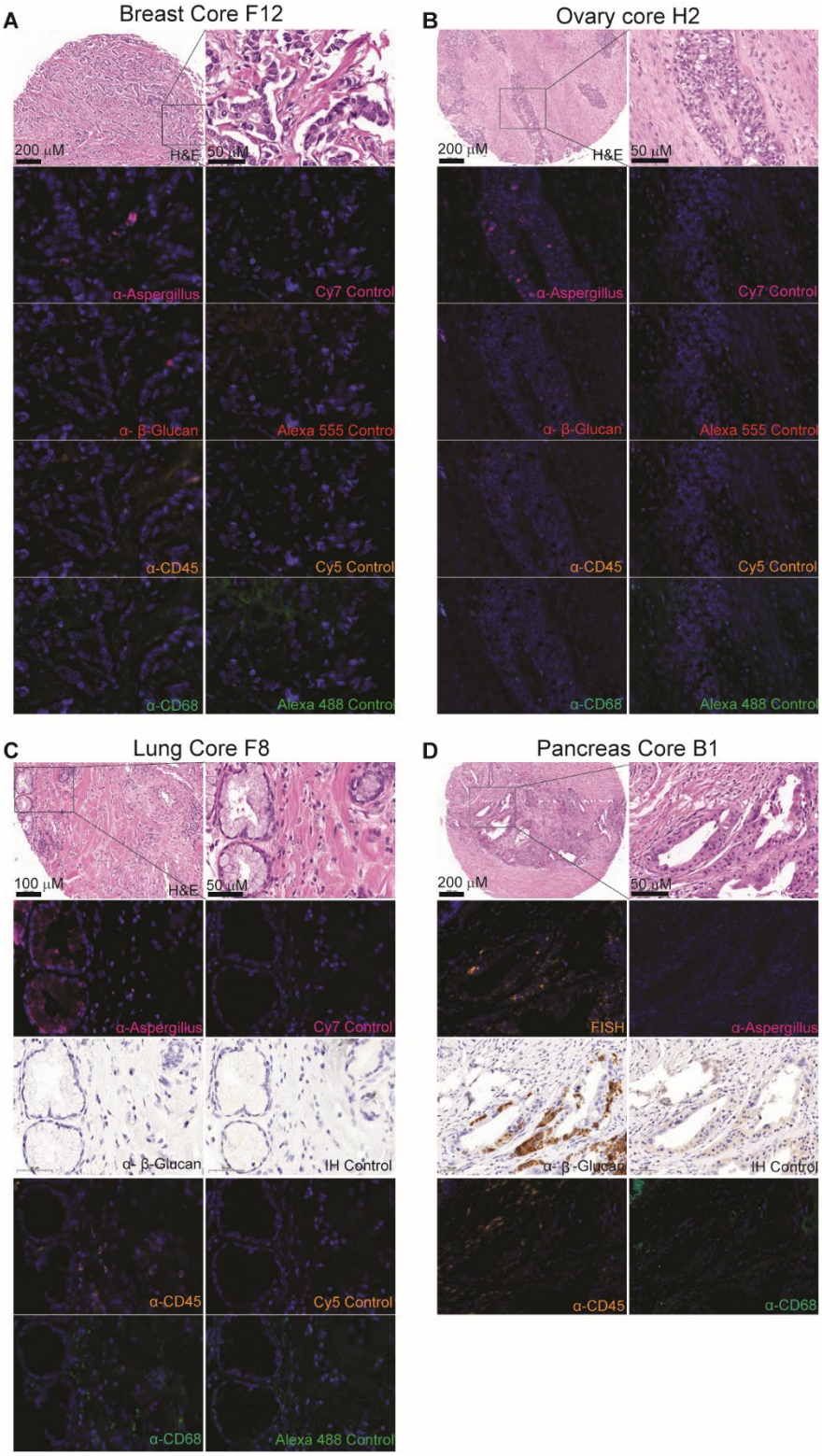


Figure 2. Visualization of fungi in human tumors. (**A-D**) Consecutive slides from representative cores from tumor microarrays of human breast cancer (**A**), ovarian cancer (**B**), lung cancer (**C**) and pancreatic adenocarcinoma (**D**) were stained with hematoxylin and eosin (H&E), antibodies against β-glucan, CD45, CD68, *Aspergillus*, or by fluorescence in situ hybridization (FISH) probes against fungal 28*S* rRNA (see Methods). Slides were also stained with only secondary antibodies as a negative control. Note that in (**D**) the core used to evaluate fluorescence negative control is missing from this slide. Scale bar for low magnification: 200μm. Square demarcates the area presented at higher magnification. Scale bar for higher magnification: 50μm.

A

Cancer type	β glucan (IF)	Aspergillus	FISH
Melanoma	12/55 (21.8%)	6/80 (8%)	0
PDAC	15/80 (18.7%)	11/70 (15.7%)	10/83 (12%)
Lung	0	7/79 (8.9%)	0
Ovary	0	20/79 (25.3%)	2/80 (2.5%)
Breast	1/99 (1%)	38/115 (33%)	0

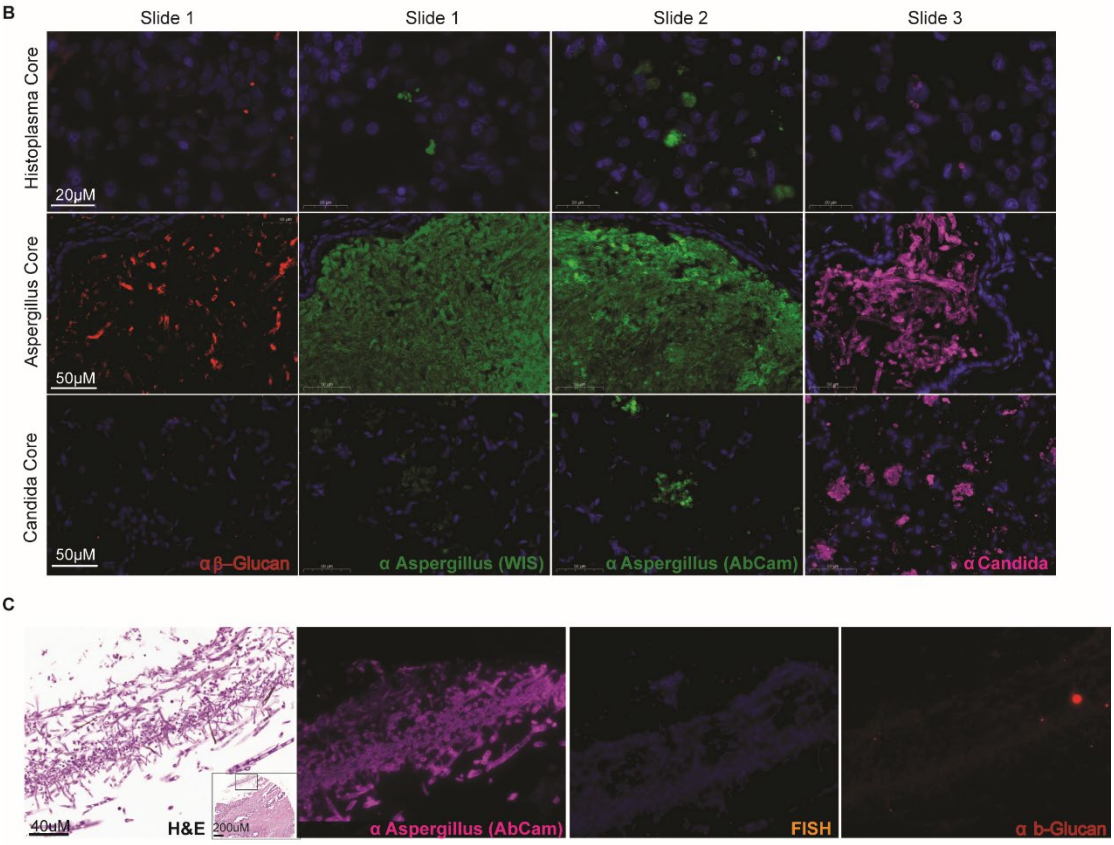


Figure 3. Fungal staining rates and validation of fungal staining methods on positive control slides. **(A)** Table summarizing positive fungal staining rates (percent of cores with positive staining) of anti- β-glucan, anti-*Aspergillus* and FISH probes in the tumor microarrays from five cancer tissue types. (**B**) FFPE slides from tissues infected with *Histoplasma* (upper panel), *Aspergillus* (middle panel) and *Candida* (lower panel), were stained with antibodies against β-glucan, *Aspergillus* (two antibodies), and *Candida*. Slide 1 was stained with antibodies against both β-glucan (red) and *Aspergillus* (green); slides 2 and 3 were stained with *anti-Aspergillus* and anti-*Candida*, respectively. Upper panel scale bar 20μm, middle and lower panels scale bar 50μm. (**C**) Consecutive slides from a FFPE tumor block that was found to be contaminated with fungi in the paraffin (and not the tissue) were stained with hematoxylin and eosin (H&E), antibodies against *Aspergillus* and β-glucan, or with fluorescence in situ hybridization (FISH) using probes against fungal 28*S* rRNA. The hyphae were only detected by the anti-*Aspergillus* antibody.

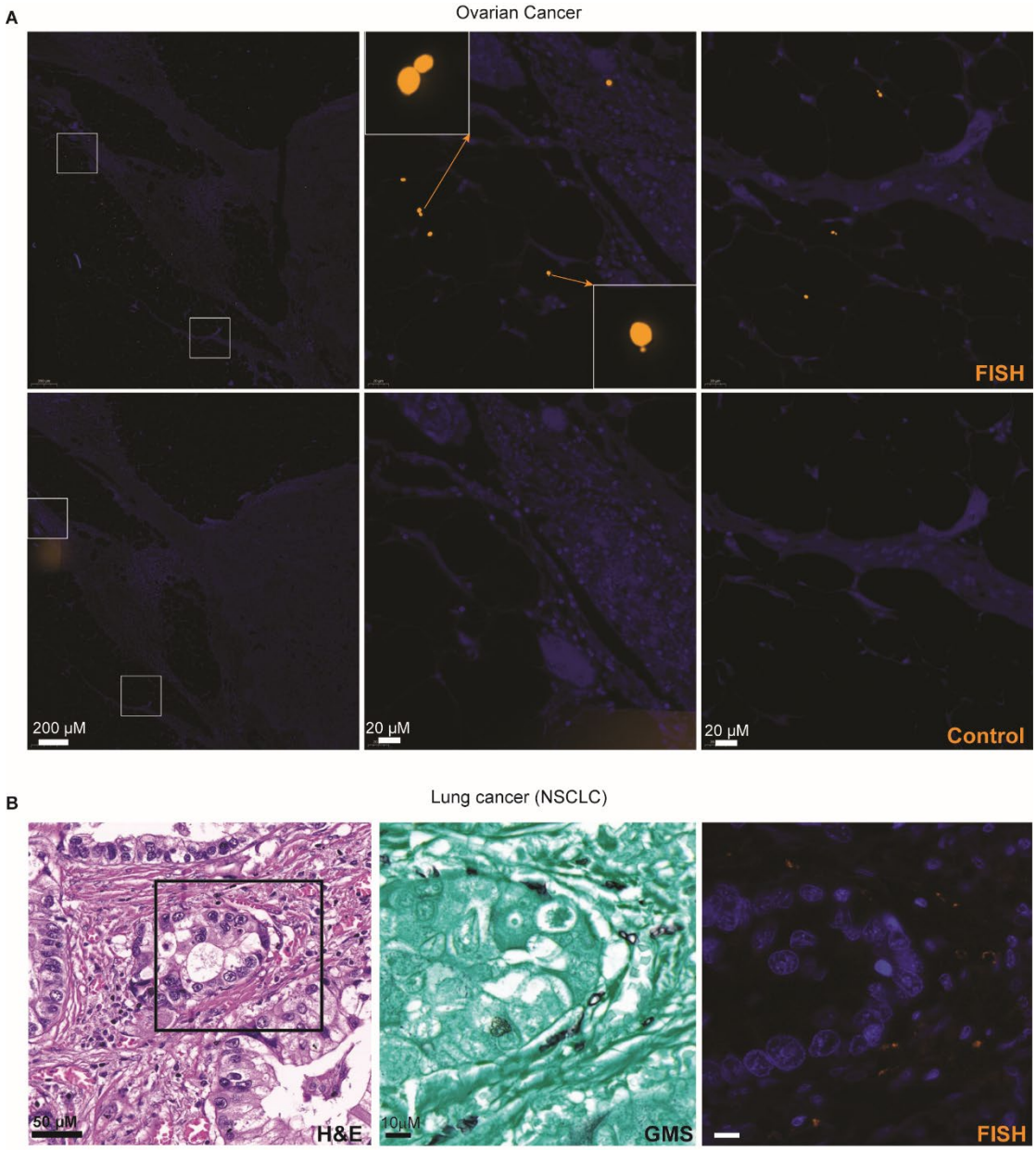


Figure 4. Detection of canonical fungal cells in tumors. (**A**) Fluorescence in situ hybridization (FISH) of a human ovarian tumor in FFPE block using 3 probes against fungal 28*S* rRNA (upper panel) or 3 scrambled probes (lower panel) (see Methods). ITS2 sequencing identified *Vishniacozyma victoriae* in this tissue sample. Scale bar in the first column is 200µm. Squares demarcates the areas presented at higher magnification in the next two columns in which the scale bar is 20µm. (**B**) A human lung tumor in FFPE block was stained with hematoxylin and eosin (H&E), Gomori methenamine silver stain (GMS), or fluorescence in situ hybridization (FISH) using the same probes as in (**A**). ITS2 sequencing identified *Fusarium keratoplasticum* and *Aspergillus tardicrescens* in this tissue sample. Scale bar in the first column is 50µm. Square demarcates the area presented at higher magnification in next two columns in which scale bar is 10µm.

Fungal nucleic acids exist in many human cancer types

Encouraged by our staining results we continued to explore fungal presence in human tumors by qPCR and NGS.

Our cohort (named Weizmann (WIS)) comprised 1,183 samples, previously examined for bacteria (*17*), of tumor, normal adjacent tissue (NAT; often paired) and normal tissue from eight tissue types (bone, breast, colon, GBM, lung, melanoma, ovary and pancreas), which were profiled for fungi using internal transcribed spacer 2 (ITS2) amplicon sequencing (Table 1, tables S1-S5). Since we expect fungal DNA in tumors to be low biomass as is in bacteria (*17*), the potential of sample contamination to overcome the signal is substantial. To account for potential contamination by environmental fungi or fungal DNA introduced during sample handling and processing, 104 paraffin-only controls and 191 DNA-extraction negative controls were included. The paraffin only controls are negative controls for center (hospital) contaminations, made by sampling paraffin only (without tissue) from a subset of the study paraffin blocks; the DNA-extraction controls are negative controls for contaminations introduced during lab processing, made by performing DNA extractions on empty tubes (with DDW only) in parallel to sample DNA extraction. These controls enabled the detection and removal of fungal contaminants and delineation of signal versus noise in the ITS2 data (See section "*Negative control samples enabled data clean up and decontamination"* and Methods). The ITS2 sequencing pipeline was rigorously validated and optimized prior to analysis. These steps will be described in detail in section "ITS2 sequencing optimization and validation in the WIS cohort" in the results.

Tissue	Sample size (# Centers)	Normal	NAT	Tumor
Breast	458 (3)	82	135	241
Lung	373 (3)	-	180	193
Melanoma	134 (3)	-	-	134
Ovary	57 (2)	-	-	57
Colon	44 (1)	-	22	22
GBM	40 (2)	-	-	40
Bone	39 (2)	-	-	39
PDAC	38 (1)	-	-	38
Total			1183	
DNA extraction controls			191	
Paraffin controls			104 (4)	

Table 1. Detailed breakdown of samples in the WIS cohort

To quantify fungal DNA load in our cohort (WIS), we used quantitative polymerase chain reaction (qPCR) of the conserved fungal 5.8S ribosomal gene in a random subset comprising 230 tumor samples and 102 negative controls. We found that all tumor types tested had on average a higher fungal load than negative controls, and fungal load differed among tumor types (Figure 5A). Fungal and bacterial load correlated across tumor types (Figure 5B), with GBM deviating from this correlation. In addition, breast cancer samples were highest in both fungal (Figure 5A) and bacterial (*17*) DNA load, suggesting tumors are polymicrobial, potentially indicating that they are more permissive to microbes and that the intratumoral microbial interactions may be more mutualistic in nature than competitive.

We then subjected all of our cohorts samples to ITS2 amplification and sequencing to characterize fungi. This analysis also found more fungal reads in samples from all cancer types than in negative controls (Figure 5C), detecting intratumoral fungi in all of the eight major human cancer types that were studied. In addition, fungal load (qPCR results) and the number of fungal reads per sample significantly correlated in tumor samples but not in the negative control samples (Figure 5D).

In collaboration with the Knight lab, we used a second cohort that encompassed whole genome sequencing (WGS) and whole-transcriptome sequencing (RNA-Seq) studies from The Cancer Genome Atlas (TCGA) (Table 2, table S5). In this cohort, reads that did not map to the human genome were aligned against a multi-domain database of 11,955 microbial (including 320 fungal) genomes (Methods). 15,512 samples (WGS: 4,736; RNA-Seq: 10,776) passed quality control with non-zero feature counts for any microbial taxa, of which 14,495 (93%) contained fungal reads. Of 6.06×10^{12} total reads in these samples, 7.13% did not map to the human genome, and 98.5% of these unmapped reads mapped to no organism in our microbial database. Of the remaining 1.5% of non-human reads that mapped to our microbial database (0.11% of total reads), 88.1% (0.097% of total) were classified as bacterial, and 2.8% (0.0031% of total) were classified as fungal, providing 1.23×10^{8} fungal reads for downstream analyses. Although contamination controls were not included in the TCGA cohort, we implemented an *in-silico* decontamination based on sequencing plate and center (*18*), and cross-referenced all fungal species against the WIS cohort, the Human Microbiome Project (HMP)'s gut mycobiome cohort (*58*), and >100 other publications to obtain a final decontaminated list (table S6) (Methods).

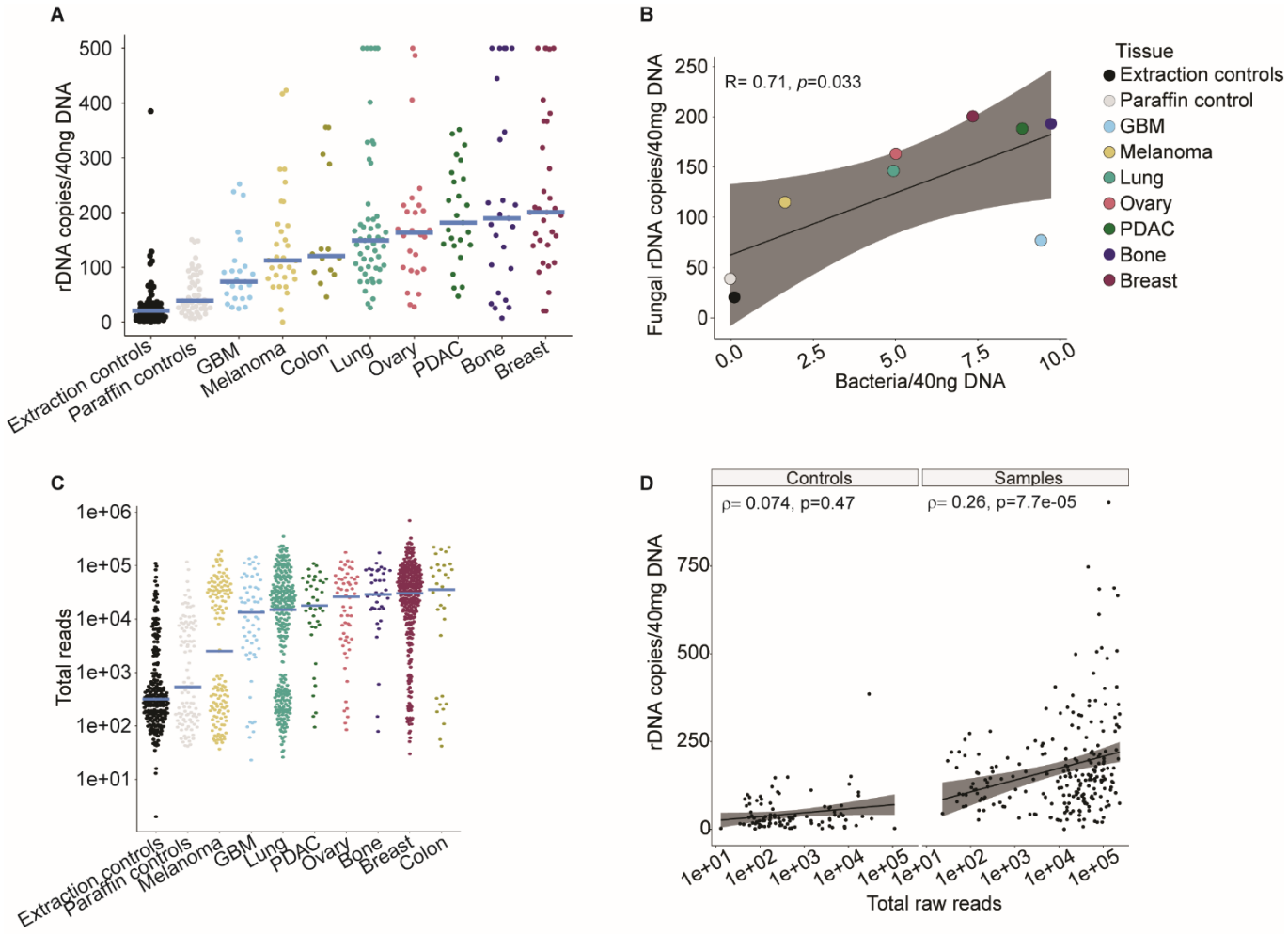


Figure 5. Fungal nucleic acids exist in human cancers. (**A**) Fungal DNA abundance in WIS cohort quantified by 5.8*S* qPCR (Methods). Blue bars show medians. Values clipped at 500. *P-values* from one-sided T-tests between tumor types and extraction controls (n=54, far left): paraffin controls (n=48), 0.04; GBM (n=25), 2.6×10^{-4}; melanoma (n=31), 9.8×10^{-7}; colon (n=14), 3.3×10^{-4}; lung (n=51), 5.2×10^{-6}; ovary (n=26), 9.2×10^{-6}; pancreas (n=25), 7.9×10^{-10}; bone (n=25), 0.02; and breast (n=33), 7.5×10^{-5}. (**B**) Scatter plot demonstrating the Pearson correlation between fungal and bacterial load as measured by qPCR across seven tumor types and controls from the WIS cohort. Fungal load is represented as rDNA copies since fungi contain a wide range of rDNA copies per cell across different species. Regression lines and confidence intervals are shown. Pearson correlation coefficient (R) and *P-value* (p) are presented. (**C**) Violin dot plot of the number of total ITS2 fungal reads (before flooring and normalization) per sample. Blue bars represent the median. (**D**) Scatter plot demonstrating the Spearman correlation between fungal load as measured by qPCR and the number of total ITS2 fungal reads sequenced across samples and controls from the WIS cohort. Regression lines and confidence intervals are shown. Spearman correlation coefficient (ρ) and *P-value* (p) are presented.

Cohort	Method	Sample types	Total # cancers	Total # samples
Weizmann (WIS)	ITS2 amplicon sequencing	Tumor, NAT, Normal, Controls	8	1183 (+295 controls)
The Cancer Genome Atlas (TCGA)	WGS, RNA-Seq	Tumor, NAT, Blood	33	15,512
Hopkins (Cristiano et al.)	WGS	Plasma	8	537
UCSD (Poore et al.)	Shotgun (WGS)	Plasma, Controls	3	169 (+58 controls)
Total	-	-	35	17,401 (+353 controls)

Table 2. Table of all cohorts studied

Motivated by the existence of ~123 million fungal reads in TCGA, despite small per-sample counts, we calculated aggregate fungal genome coverage across all WGS and RNA-Seq samples (table S7; Methods). This revealed 31 fungi with ≥1% genome coverage and several with high or nearly complete coverage, including *Saccharomyces cerevisiae* (99.7% coverage), *Malassezia restricta* (98.6% coverage), *Candida albicans* (84.1% coverage), *Malassezia globosa* (40.5% coverage), and *Blastomyces gilchristii* (35.0% coverage). Fungal species from the WIS cohort that overlapped with TCGA were significantly more likely to have ≥1% genome coverage than non-WIS-overlapping species (Fisher exact test: $p=1.05\times10^{-8}$, odds ratio=13.1).

In the TCGA cohort, we observed in 31 of 32 cancer types that the proportions of bacterial reads in primary tumors were significantly higher than fungal reads (Figure 6A), and all cancer types had significantly higher bacterial proportions after normalizing by genome sizes (data not shown). Calculating average relative abundances among the bacterial and fungal data in TCGA primary tumors revealed 86.7% bacterial and 13.3% fungal without normalizing by respective genome sizes, or 96% bacterial and 4% fungal after normalizing by genome sizes, suggesting that the bacteriome constitutes most of the tumor microbiome. Fungal and bacterial read proportions had high Spearman correlations (Figure 6B), including primary tumors, normal adjacent tissues (NATs), and blood. These data also support a bacterial-dominated but inherently polymicrobial cancer microbiome.

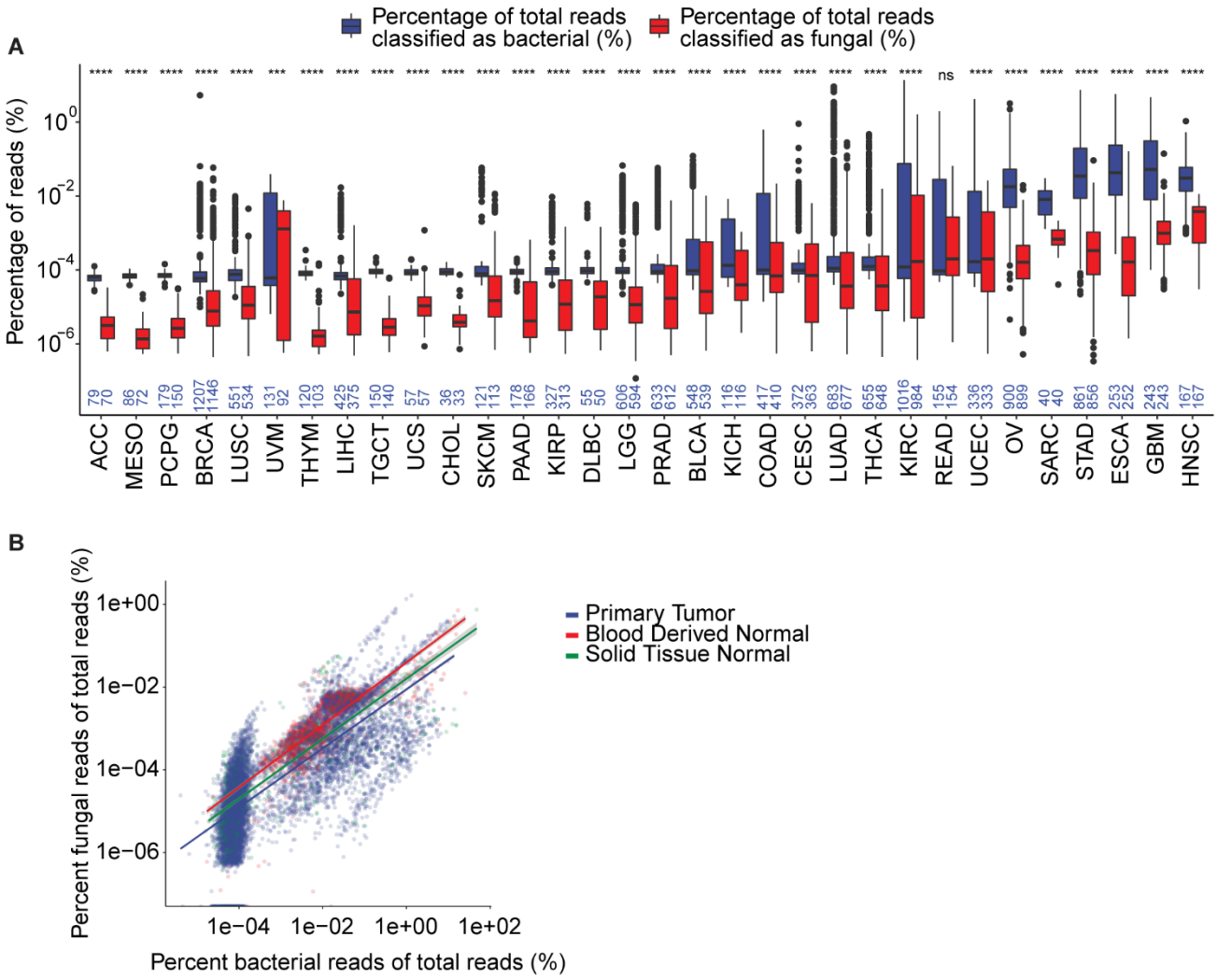


Figure 6. Fungal nucleic acids exist in human cancers, TCGA data. (**A**) Percentage of reads in TCGA primary tumors mapped to fungal and bacterial genomes versus total reads. Sample sizes inset in blue and vary slightly when samples had non-zero bacterial counts but zero fungal counts. Two-sided Wilcoxon tests for each cancer type fungal vs. bacterial comparisons; **** indicates $p<0.001$, ns=not significant. Box plots show median line, 25th and 75th percentiles, and 1.5× interquartile range. (**B**) Log_{10}-scaled scatter plot of the percent of fungal vs. bacterial reads with respect to total reads in the concomitant bam files. Linear regression lines are overlaid on the scatter plot, colored by the respective sample type. Non-parametric Spearman correlation testing revealed significant associations between proportions of fungal and bacterial reads: Primary tumor, $\rho=0.76$, $t=1.7\times10^{8}$, $p\approx0$; blood derived normal, $\rho=0.76$, $t=6.4\times10^{10}$, $p\approx0$; solid tissue normal, $\rho=0.84$, $t=5.1\times10^{7}$, $p\approx0$.

Our cohort (WIS) and the TCGA cohort have complementary advantages and drawbacks; together they complement each other and strengthen our findings and conclusions. Advantages of the WIS cohort include aseptic sample curation and processing, using a mechanical shearing step to optimize microbial DNA extraction, inclusion of hundreds of experimental contamination controls, complementary tissue imaging, and fungal-specific qPCR, which together substantiate confidence in the true presence/absence of intratumoral fungi. However, the nature of ITS2 amplicon sequencing precludes genome-wide coverage analyses. Conversely, the shotgun metagenomic approach taken with the TCGA cohort paired with very large sample sizes enables fungal comparison with host information, detection across most human cancer types, and represents a scalable approach compatible with historical data; however, due to the lack of experimental contamination controls, it relies on *in-silico* decontamination with less confidence in presence/absence calls.

Metagenomic analyses demonstrate cancer type-specific mycobiomes

Metagenomic analysis allowed us to characterize the fungal composition within our different samples, gain insight into the fungal profile of tumor types as a whole and learn about the spread of individual fungi in the different samples. Mycobiome richness varied significantly across cancer types (Figure 7A). In addition, the richness of tumor samples in our cohort as well as the TCGA cohort was significantly lower for fungi vs. bacteria (Figure 7B), similar to previous findings in the gut microbiome (*58*). Richness of both fungi and bacteria was lower in our cohort (WIS) relative to the TCGA cohort (shotgun metagenomics), likely due to (i) numerous negative controls in the WIS cohort, (ii) flooring of the WIS data to counteract index-hopping sequencing noise (*59*), and (iii) potential read splitting during shotgun metagenomics alignments in the TCGA cohort (Methods). Interestingly, 71% (5/7) of cancer types shared by both cohorts showed significant positive correlations between intratumoral fungal and bacterial richness (Figure 7C). While intratumoral mycobiome α-diversity was low, the β-diversity was high between tumor samples (Figure 7D), preventing saturation in the rarefaction plots (Figure 7E). Nevertheless, we found that samples within cancer types clustered together by their mycobiome (data not shown; PERMANOVA: $p=0.037$). β-diversity analyses within TCGA sequencing centers similarly revealed cancer-type specific mycobiome compositions (data not shown; PERMANOVA: $p=0.001$).

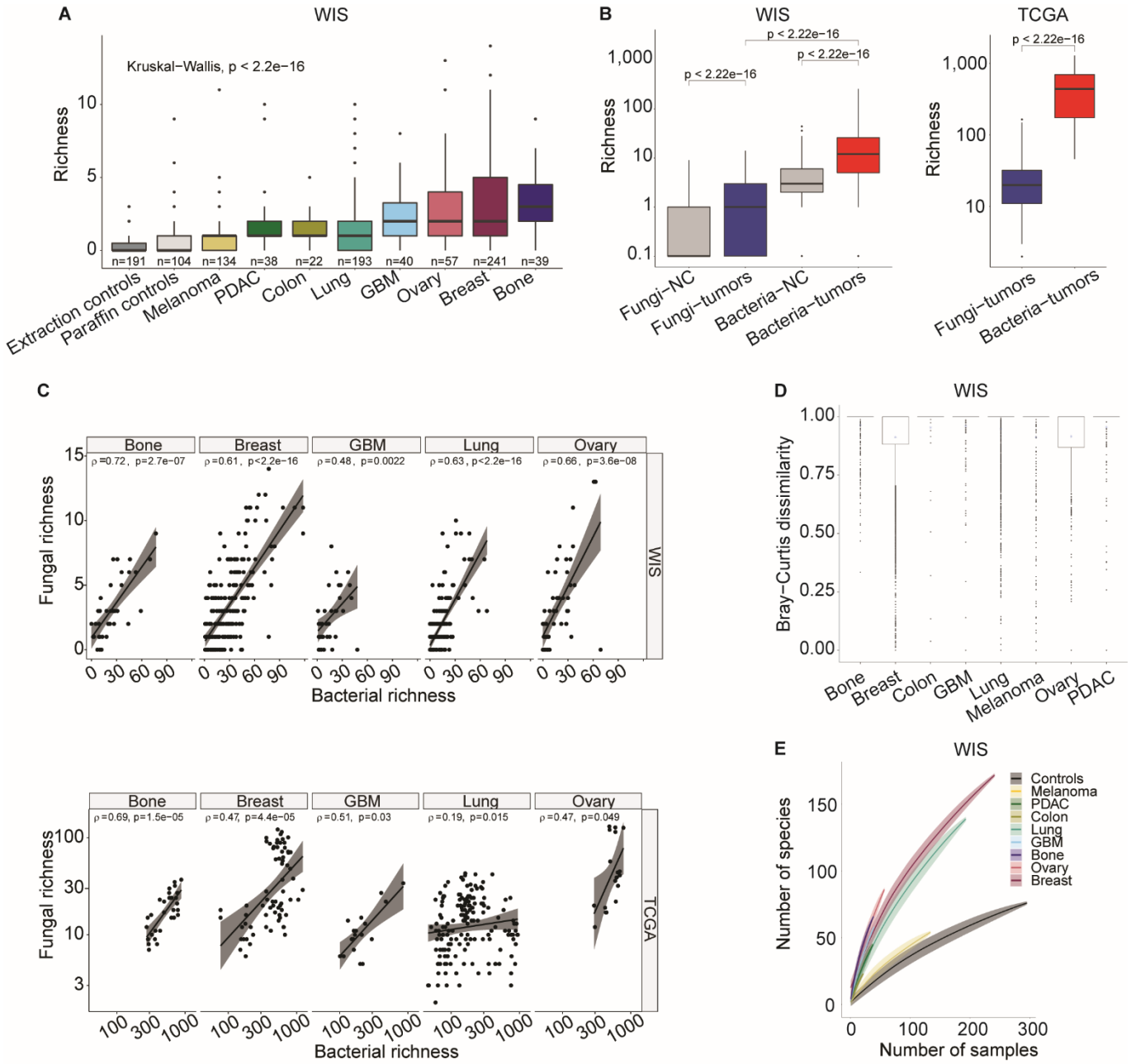


Figure 7. Mycobiome richness varies across cancer types. (**A**) Box plot of the number of species (richness) per WIS sample in different tumor types after data flooring and normalization (Methods). Only fungi that passed the filtering steps in any of the tumor types were included in the analysis. Kruskal-Wallis test reflects significant richness variation across cancer types. (**B**) Fungal and bacterial species richness for WIS and TCGA cohorts. NC: negative controls. (**C**) Scatter plot demonstrating Spearman correlations (ρ) and concomitant *P-values* between fungal and bacterial richness across five tumor types shared between the WIS and TCGA cohorts. Linear regression lines and 95% confidence intervals shown. (**D**) Box plot of Bray-Curtis dissimilarity scores within tumor type. The asterisks depict the mean. (**E**) Rarefaction plot of the number of species detected in the WIS cohort per tumor type with 100 random subsamples per number of samples. Mean and standard deviation shown. Extraction and paraffin controls were grouped together. (**A, B, D**) Box plots show median line, 25th and 75th percentiles, and 1.5× interquartile range.

Across the eight cancer types tested in the WIS cohort, Ascomycota and Basidiomycota phyla dominated the intratumoral mycobiome (Figure 8A). Ascomycota and Basidiomycota are the two main fungal phyla, together comprising over 95% of the fungal kingdom. The Ascomycota to Basidiomycota ratio (A/B ratio) was highest in colon cancer, due to abundant Saccharomycetes, and lowest in melanoma, due to abundant Malasseziomycetes. These differences correspond to known fungal taxa that inhabit the gut (*58*) and skin (*34*), suggesting partial conservation of normal tissue-specific ecologies in tumors. Indeed, unsupervised clustering of tumors alongside normal and normal adjacent tissue (NAT) samples showed tissue-specific clustering by the most prevalent fungi in these tissues by both fungal prevalence and mean relative abundance (Figure 8B-C). In addition, both WIS and TCGA cohorts demonstrated co-clustering of tumor and NAT samples when comparing β-diversity scores, supporting similar tumor and NAT compositions (Figure 8D-E). Bray-Curtis dissimilarity between tumor and NAT samples from the same patient demonstrated higher similarity vs. non-matched tumor and NAT samples or tumor sample pairs (Figure 8F). To test whether this is the cause for the co-clustering of tumor and NAT in the PCoA, we repeated the PCoA analysis without the matched tumor-NAT pairs that originate from the same patient. Co-clustering of tumor and NAT profiles still occurred after discarding from the analysis pairs of tumor-NAT samples from the same patients (Figure 8G). Collectively, these analyses portray ubiquitous, low-abundance, cancer type-specific mycobiomes that have ecologies similar to those in normal adjacent tissues.

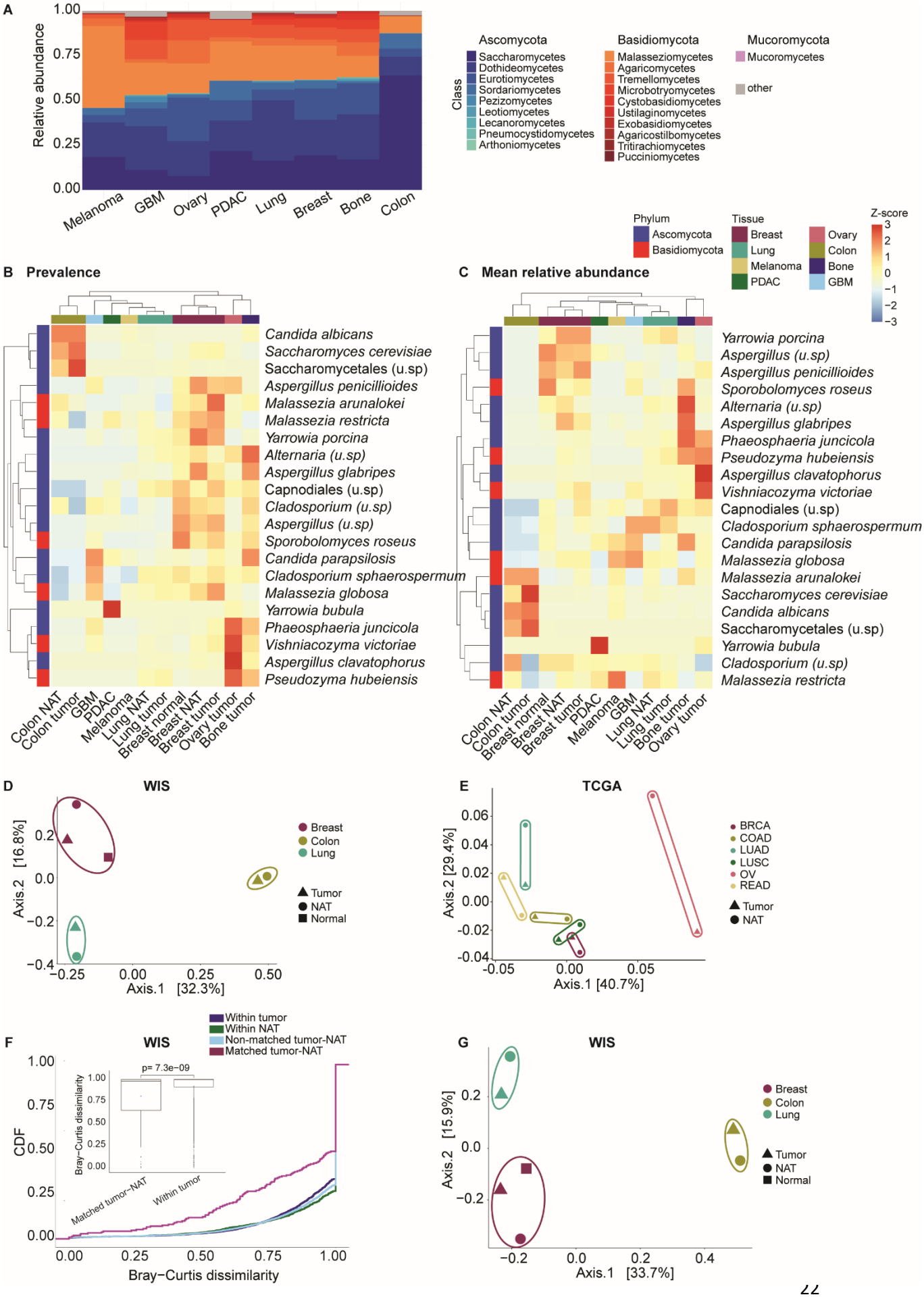

A
Relative abundance
1.00
0.75
0.50
0.25
0.00
Melanoma
GBM
Ovary
PDAC
Lung
Breast
Bone
Colon
Class
Ascomycota
Saccharomycetes
Dothideomycetes
Eurotiomycetes
Sordariomycetes
Pezizomycetes
Leotiomycetes
Lecanoromycetes
Pneumocystidomycetes
Arthoniomycetes
Basidiomycota
Malasseziomycetes
Agaricomycetes
Tremellomycetes
Microbotryomycetes
Cystobasidiomycetes
Ustilaginomycetes
Exobasidiomycetes
Agaricostilbomycetes
Tritirachiomycetes
Pucciniomycetes
Mucoromycota
Mucoromycetes
other
Phylum
Ascomycota
Basidiomycota
Tissue
Breast
Lung
Melanoma
PDAC
Ovary
Colon
Bone
GBM
Z-score
3
2
1
0
−1
−2
−3
B Prevalence
Candida albicans
Saccharomyces cerevisiae
Saccharomycetales (u.sp)
Aspergillus penicillioides
Malassezia arunalokei
Malassezia restricta
Yarrowia porcina
Alternaria (u.sp)
Aspergillus glabripes
Capnodiales (u.sp)
Cladosporium (u.sp)
Aspergillus (u.sp)
Sporobolomyces roseus
Candida parapsilosis
Cladosporium sphaerospermum
Malassezia globosa
Yarrowia bubula
Phaeosphaeria juncicola
Vishniacozyma victoriae
Aspergillus clavatophorus
Pseudozyma hubeiensis
Colon NAT
Colon tumor
GBM
PDAC
Melanoma
Lung NAT
Lung tumor
Breast normal
Breast NAT
Breast tumor
Ovary tumor
Bone tumor
C Mean relative abundance
Yarrowia porcina
Aspergillus (u.sp)
Aspergillus penicillioides
Sporobolomyces roseus
Alternaria (u.sp)
Aspergillus glabripes
Phaeosphaeria juncicola
Pseudozyma hubeiensis
Aspergillus clavatophorus
Vishniacozyma victoriae
Capnodiales (u.sp)
Cladosporium sphaerospermum
Candida parapsilosis
Malassezia globosa
Malassezia arunalokei
Saccharomyces cerevisiae
Candida albicans
Saccharomycetales (u.sp)
Yarrowia bubula
Cladosporium (u.sp)
Malassezia restricta
Colon NAT
Colon tumor
Breast normal
Breast NAT
Breast tumor
PDAC
Melanoma
GBM
Lung NAT
Lung tumor
Bone tumor
Ovary tumor
D
WIS
Axis.2 [16.8%]
0.2
0.0
−0.2
−0.4
−0.25
0.00
0.25
0.50
Axis.1 [32.3%]
Breast
Colon
Lung
Tumor
NAT
Normal
E
TCGA
Axis.2 [29.4%]
0.06
0.04
0.02
0.00
−0.02
−0.04
-0.05
0.00
0.05
Axis.1 [40.7%]
BRCA
COAD
LUAD
LUSC
OV
READ
Tumor
NAT
F
WIS
Within tumor
Within NAT
Non-matched tumor-NAT
Matched tumor-NAT
CDF
1.00
0.75
0.50
0.25
0.00
p= 7.3e−09
Bray−Curtis dissimilarity
1.00
0.75
0.50
0.25
0.00
Matched tumor−NAT
Within tumor
0.00
0.25
0.50
0.75
1.00
Bray−Curtis dissimilarity
G
WIS
Axis.2 [15.9%]
0.2
0.0
−0.2
−0.2
0.0
0.2
0.4
Axis.1 [33.7%]
Breast
Colon
Lung
Tumor
NAT
Normal

Figure 8. Metagenomics analyses demonstrate cancer type-specific mycobiomes. (**A**) Mean relative abundance bar plots at class-level phylotypes across WIS cohort tumor types. Colors correspond to the phyla of each class. Blues: Ascomycota; reds: Basidiomycota; pink: Mucoromycota. (**B, C**) Unsupervised hierarchical clustering of fungal (**B**) prevalence or (**C**) mean relative abundance in the WIS cohort using species that appear in ≥10% of samples in ≥1 condition. Values represent Z-scores per row. u.sp: unknown species. (**D**) Principal coordinate analysis (PCoA) on the Jaccard dissimilarity indexes using fungal species across tissues in WIS cohort. (**E**) Bray-Curtis PCoA of averaged relative abundances on rescaled, Voom-SNM corrected TCGA WGS and RNA-Seq data (see Methods) on cancer types also found in the WIS cohort and with at least 10 tumors and NATs available in TCGA. Sample counts: breast NAT, n=100; breast tumor, n=978; colorectal NAT, n=72; colorectal tumor, n=526; Lung NAT, n=194; Lung tumor, n=1068; Ovarian NAT, n=10; Ovarian tumor, n=683. Note that "lung" combines TCGA projects LUAD and LUSC and that "colorectal" combines TCGA projects COAD and READ. (**F**) Cumulative Distribution Function (CDF) plot of Bray-Curtis dissimilarity scores in the WIS cohort, within tumor samples, within NAT samples, between tumor and NAT samples from different patients ("Non-matched tumor-NAT"), and between paired tumor and NAT samples from the same patient ("Matched tumor-NAT"). All scores of pairs were calculated within a tumor type and included only tissue types for which NAT samples were available: breast, lung and colon. Inset: Boxplot of dissimilarity measurements between the matched samples group and within tumor samples. (**G**) PCoA on the Jaccard dissimilarity indices between species profiles of the different tissue types after discarding paired tumor-NAT patient samples in WIS cohort. This analysis was done to demonstrate that the tumor-NAT clustering that was observed is not the result of high similarity between mycobiomes of samples that originate from the same patients. We have thus removed from the analysis either the tumor or the NAT samples (by random) from patients that had both sample types.

Machine learning analyses demonstrated cancer-type specific mycobiomes

Since we detected tissue type specific mycobiome profiles in both WIS and TCGA data sets we attempted to build a pan-cancer classifier using these profiles. We combined mycobiome data from all TCGA sequencing centers and experimental strategies using supervised batch correction on 14,495 samples, as previously done with TCGA bacteriomes and viromes (*18*) (Methods). Evaluating one-cancer-type-versus-all-others models on the normalized mycobiome data provided strong discriminatory performance across 32 cancer types (Figure 9A). In addition, scrambled and shuffled negative controls showed null, significantly worse performance for every cancer type (Figure 9A). Machine learning (ML) was done by using three feature groups: Fungi in the TCGA cohort with high coverage (>1% of genome), fungal species in the TCGA cohort overlapping to fungal species detected in our cohort (WIS) and all fungal features detected in the TCGA cohort after decontamination. The feature group of fungi intersecting with our cohort (WIS) gave results comparable to both other feature groups, suggesting that the fungal species detected in our cohort represent a core, clean feature

set generalizable to other cohorts. We validated this by testing the performance of equal sized feature groups that do not appear in our cohort in ML (Figure 9B), these groups showed significantly lower AUROC scores relative to the WIS overlapping feature set. Furthermore, combining fungal and bacterial information from our cohort revealed synergistic performance benefits (Figure 9C). We then independently batch corrected two stratified TCGA halves, separately trained ML models on them, followed by testing on each opposing half—this revealed significantly correlated performance between primary tumor comparisons (data not shown), suggesting generalizable discriminatory performance.

Evaluating one-cancer-type-versus-all-others models on our data set (WIS) also provided discriminatory performance in all cancer types with synergistic performance of fungal and bacterial features (Figure 10A).

Previous bacteriome-centric analyses revealed cancer type-specific, blood-derived microbial DNA (*18*), prompting us to examine fungal DNA in TCGA WGS blood samples. Indeed, evaluating models on the batch-corrected dataset showed strong pan-cancer discriminatory performance (Figure 10B). All controls performed for the tumor tissue above were repeated for the blood analysis with similar results (data not shown). Collectively, these ML analyses support cancer-type specific tissue and blood mycobiomes, which may have clinical utility.

Compositional similarity between tumor and NAT mycobiome samples (Figure 8B-G) indicated that their discrimination may be challenging. Indeed, ML on most TCGA raw data subsets (data not shown) and on the WIS data demonstrated weak performance (Figure 11A). Nonetheless, the small average effect size between tumor and NAT seemed surmountable when re-examining the full, batch corrected TCGA dataset (Figure 11B). Affirming the issue of effect size, comparing breast tumors to true normal tissue in the WIS cohort revealed more differential fungal prevalences and better ML performance (Figure 11C-D). These analyses suggest that tissue mycobiomes may distinguish tumor and NAT if studies are sufficiently powered.

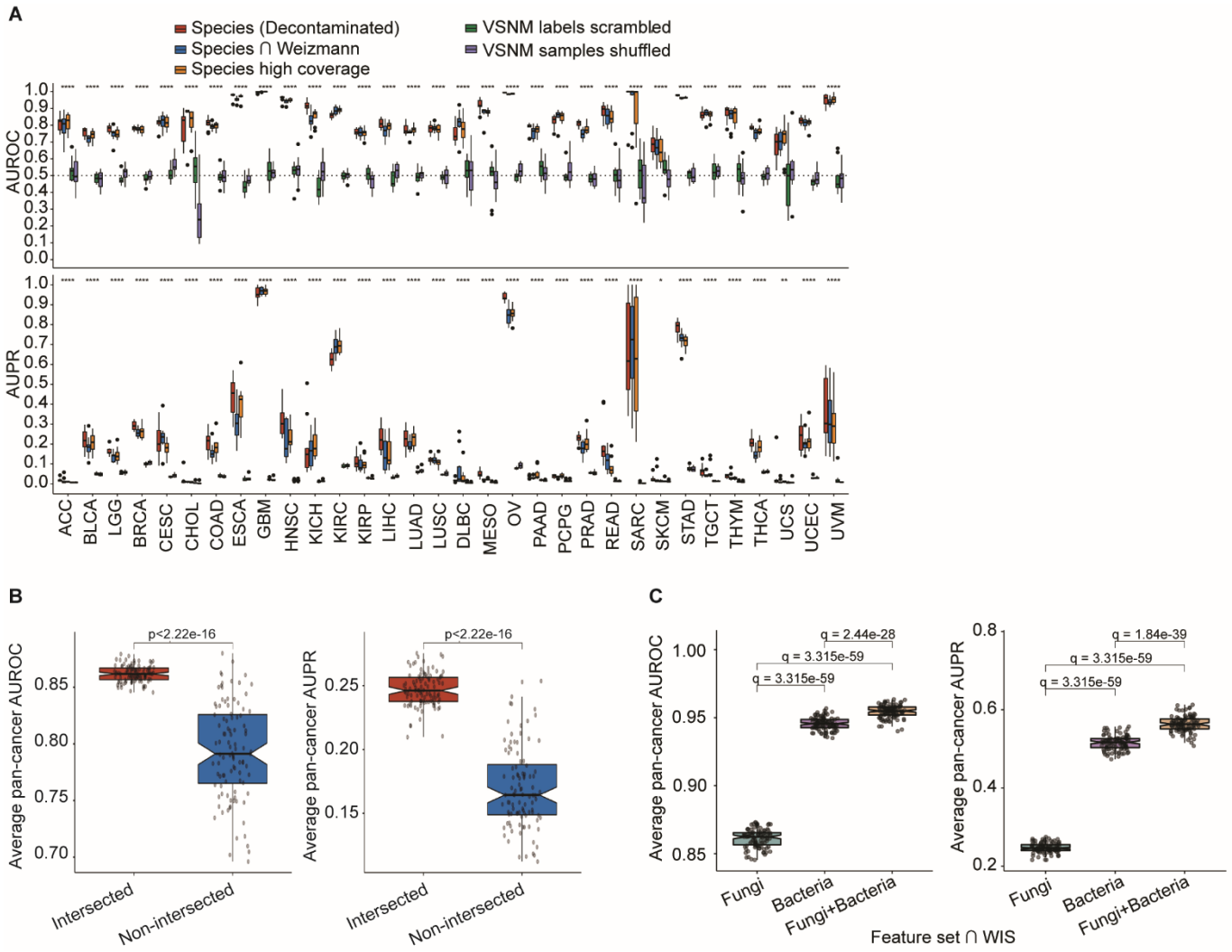


Figure 9. Machine learning (ML) analyses reveal cancer type-specific mycobiomes in tumor tissues. (**A**) One-cancer-type-versus-all-others predictions using batch-corrected, TCGA primary tumor data (n=10,998). Scrambled and shuffled machine learning negative controls were repeated on pan-cancer, batch-corrected primary tumor data and compared to performance using biological samples, which included 224 decontaminated fungal species, 34 WIS-overlapping fungal species, or 31 fungal species with ≥1% aggregate coverage. For hypothesis testing, biological data and scrambled/shuffled controls were aggregated into two separate groups, and two-sided Wilcox tests were applied per cancer type per performance metric (AUROC or AUPR). *: p≤0.05; **: p≤0.01; ***: p≤0.001; ****: p≤0.0001. (**B**) To test whether the 34 WIS-overlapping species provided greater discriminatory performance in TCGA than other detected fungi, multi-class machine learning models were built on stratified splits of WGS samples in TCGA with 70% training and 30% holdout test sets (Methods) using only the 34 WIS-overlapping species or 34 non-WIS-overlapping randomly selected fungi. This process was repeated for 100 iterations, and AUROC (left) and AUPR (right) performance was calculated on the 30% holdout test set for each iteration. Two-sided Wilcox tests *P-values* are presented. (**C**) Multi-class pan-cancer discrimination among TCGA WGS tumor samples using WIS-overlapping features across 100 iterations of stratified train-test splits (see Methods).

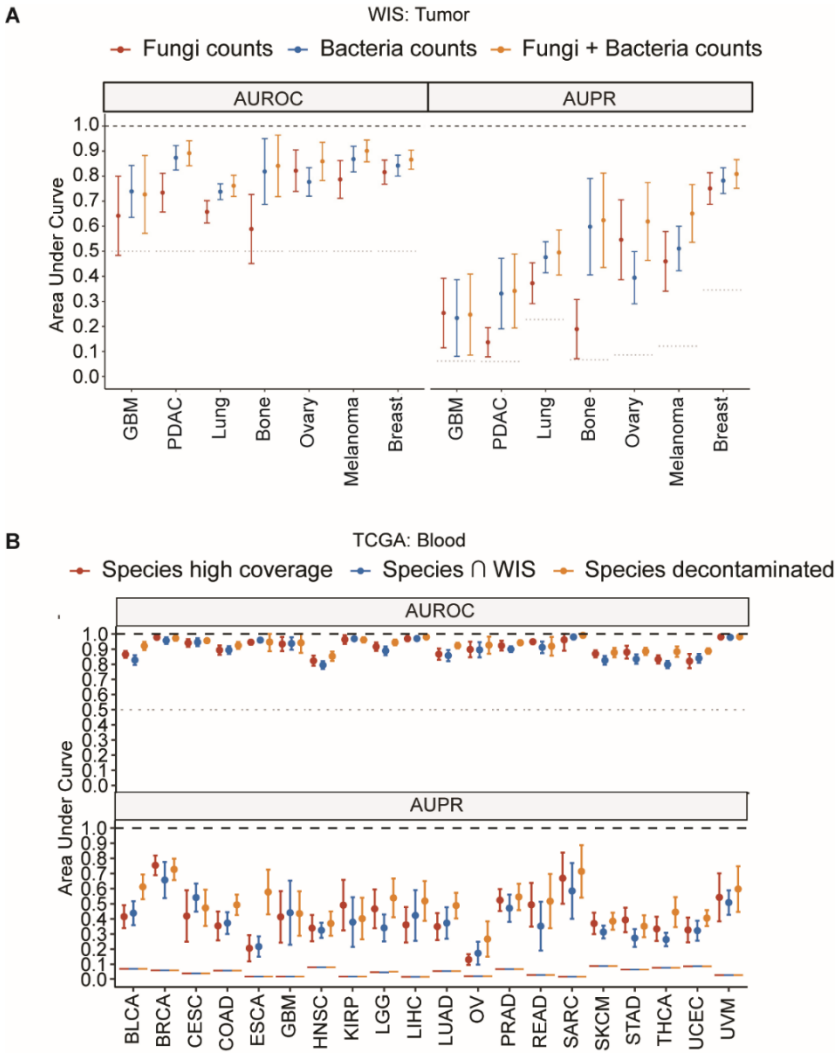


Figure 10 . Machine learning (ML) analyses reveal cancer type-specific mycobiomes in WIS tumor tissues and in TCGA blood samples. (**A**) Ten-fold cross-validation ML approach was applied to the WIS cohort data, using fungal, bacterial, or fungal and bacterial raw counts to discriminate one cancer type versus all others. All filtered fungal hits across all taxa levels ("free rank") were included. Dots denote average performance and error bars denote 95% confidence intervals. Gray horizontal dots denote the null AUROC and AUPR values, the latter of which is the prevalence of the positive class (here, each cancer type). (**B**) One-cancer-type-versus-all-others predictions using batch-corrected, TCGA blood data (n=1771). "High coverage," 31 fungal species with ≥1% aggregate genome coverage; "Species ∩ WIS," 34 WIS-overlapping fungal species; "decontaminated," 224 decontaminated fungal species. Horizontal lines denote null AUROC or AUPR.

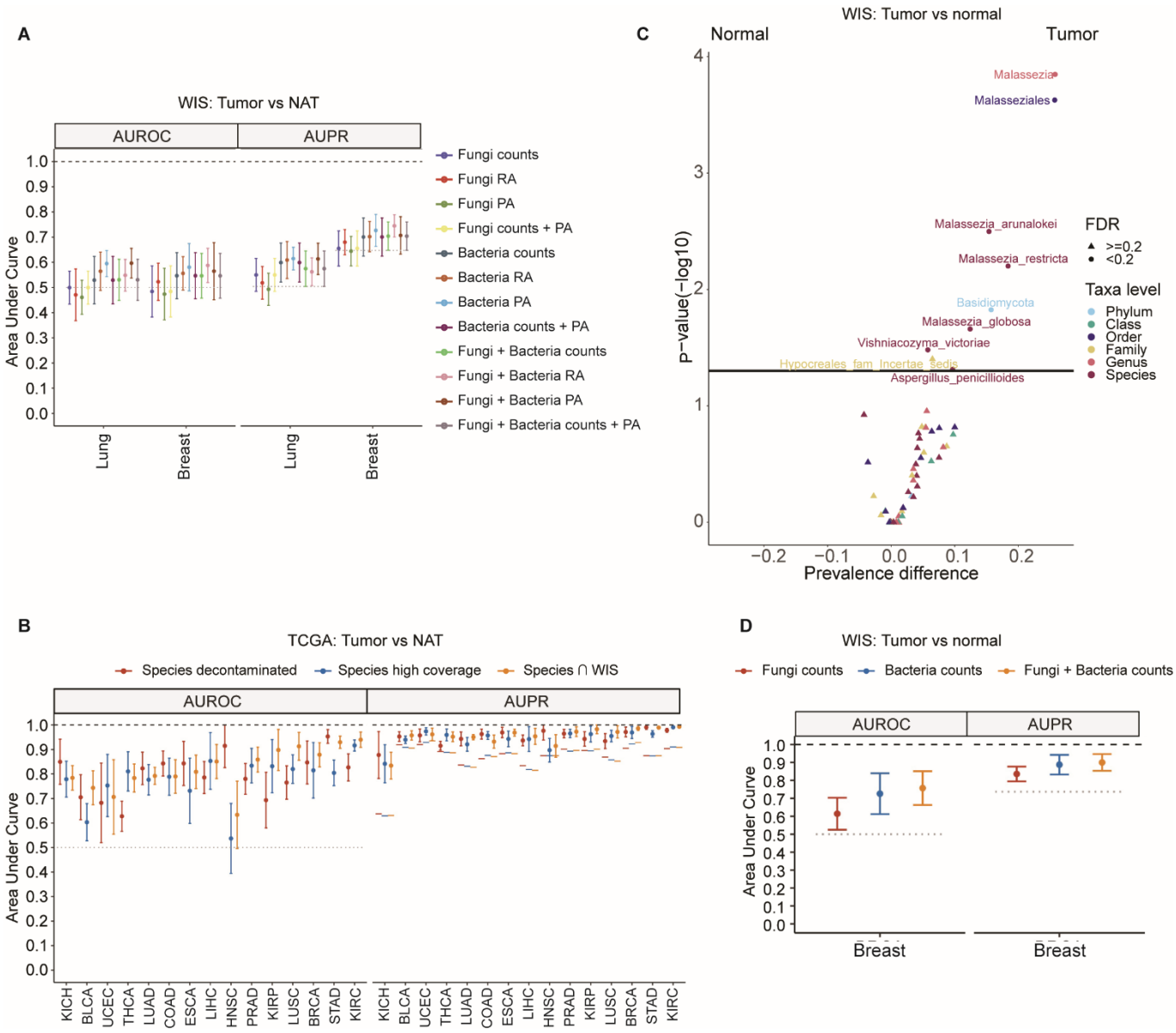


Figure 11. Machine learning (ML) analyses for tumor versus NAT discrimination necessitates a large sample size. (**A**) Ten-fold cross-validation machine learning models tested to discriminate tumor versus NAT samples using the WIS cohort fungal, bacterial, or fungal and bacterial data comprising raw counts, relative abundances, binary presence-absence, or combined counts and presence-absence information. Microbial hits at all taxa levels that passed filtering were included ("free rank"). At least 20 samples in each class were required to be tested. (**B**) Ten-fold cross-validation machine learning models tested to discriminate TCGA tumor vs. NAT samples using pan-cancer batch corrected data. Feature subsets included 224 decontaminated fungal species (red), 31 fungal species with ≥1% aggregate coverage ("high coverage"), and 34 fungal species that overlapped with the WIS cohort ("∩ Weizmann (**C**) Differential prevalence testing in the WIS cohort between breast cancer tumor samples and true normal breast tissue samples across all taxa levels. Colors represent taxa level and shapes represent FDR cutoff. (**D**) Ten-fold cross-validation ML models built to discriminate WIS breast cancer tumor samples vs. true normal breast tissue based on fungi, bacteria, or fungi and bacteria raw counts. Microbial hits at all taxa levels that passed filtering were included ("free rank"). (**A, B, D**) Dots denote average values and error bars denote 95% confidence intervals. Horizontal gray/colorful dots denote the null AUROC and AUPR values. Null AUPR values vary slightly when subsetting feature sets resulted in zero sum samples that had to be removed prior to batch correction and machine learning.

Blood mycobiome profiles discriminate between cancer patients and healthy individuals

Blood-derived cancer type specific fungal compositions in TCGA suggest their utility as minimally invasive diagnostics, analogous to bacterial counterparts (*18*). We thus sought to validate these findings in two independent, previously published cohorts (Hopkins, UCSD), together comprising 330 healthy and 376 cancer-bearing subjects (table S5), that underwent low-coverage whole genome plasma sequencing (*18*, *60*). Notably, the Hopkins cohort concentrated on treatment-naive, early-stage cancers while the UCSD cohort focused on treated, late-stage cancers, collectively addressing most clinical scenarios across 10 cancer types.

In the Hopkins cohort, decontaminated fungal species (n=209) provided moderate discriminatory performance, and performance with multi-domain feature sets exceeded published host-centric approaches including when subsetting to 287 fungal and bacterial species overlapping to features from our cohort (WIS) (Figure 12A). ML of individual cancer types vs. controls performed similarly (Figure 12B), with the best fungal performance in breast cancer. Testing ML models in a one-cancer-type-versus-all-others manner similarly revealed moderate discrimination for decontaminated fungi and strongest discrimination with multi-domain features (Figure 12C). ML across individual stages vs. healthy continued this pattern for all stages (Figure 12D), suggesting that microbial-augmented liquid biopsies are not dependent on cancer stage. UCSD cohort analysis showed similar results (data not shown). Collectively, these analyses suggest the clinical utility of multi-domain microbial nucleic acids in plasma samples from treatment-naive patients. Furthermore, seeing that the feature set of bacteria and fungi that overlaps with the WIS cohort provided nearly equivalent discriminatory performance as a multi-domain database 26-fold larger, suggests a significant tumor origin of the species from the WIS cohort, generalizable across additional cohorts.

Cancer mycobiome components are associated with patients clinical parameters

We next explored the diagnostic and prognostic capacities of the cancer mycobiome, which were previously established for cancer bacteriomes (*17*, *18*, *61*). We tested whether disease phenotypes, patient survival, and treatment response were associated with fungal biomarkers.

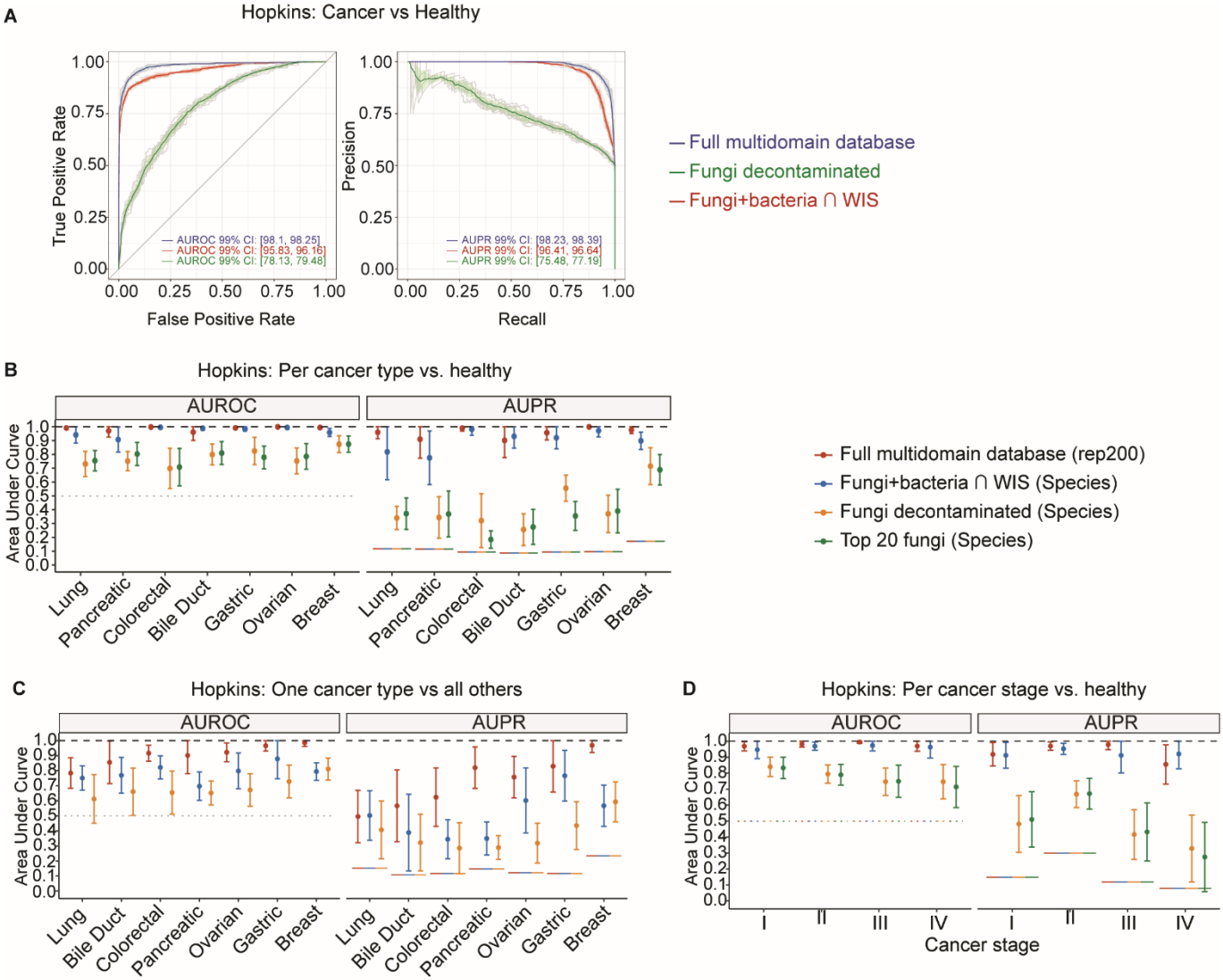
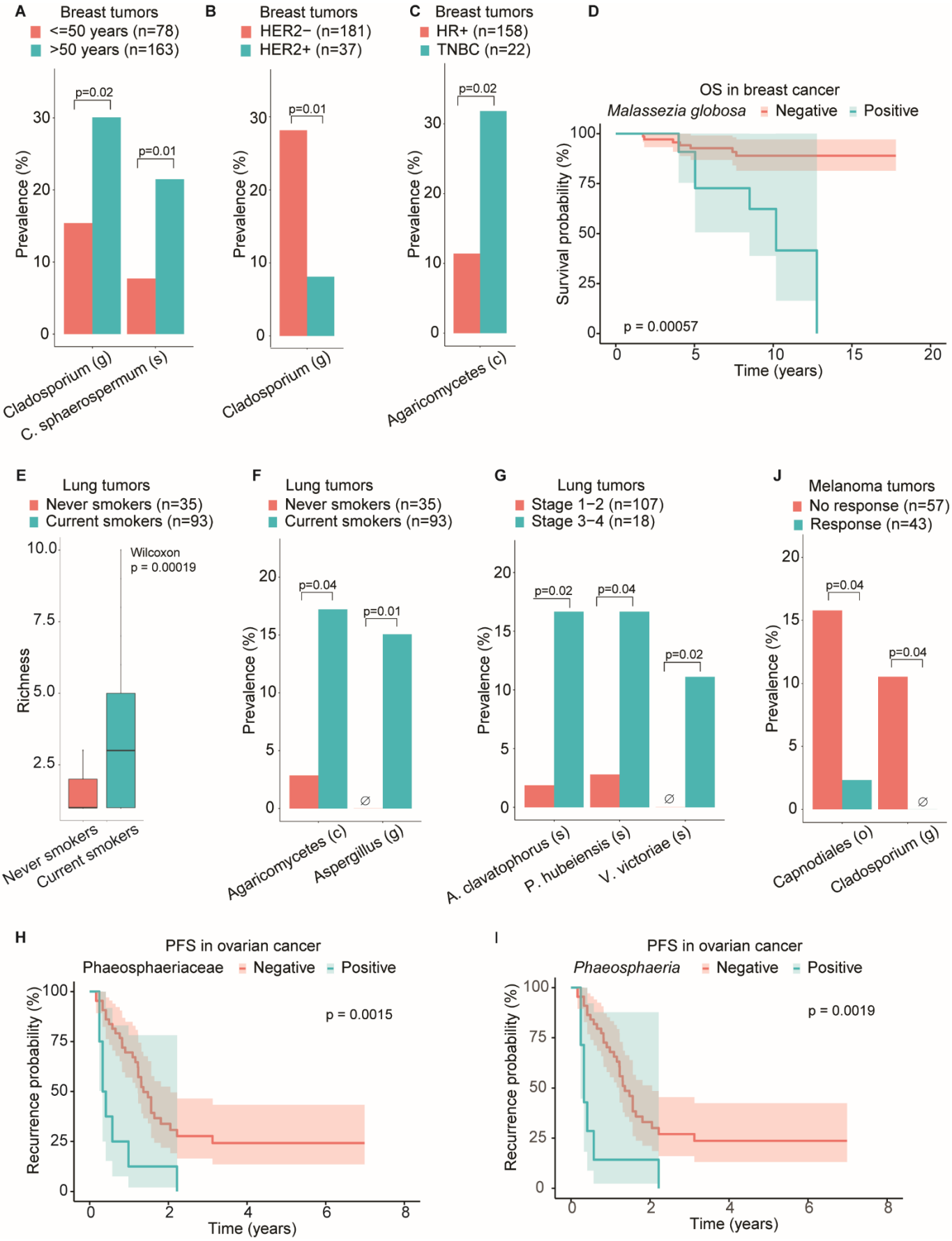


Figure 12. Blood mycobiome profiles discriminate between cancer patients and healthy individuals. (**A**) Treatment-naive pan-cancer vs. healthy discrimination in the Hopkins plasma cohort using ten-fold cross-validation repeated ten times across all database hits (blue, 7418 features), WIS-overlapping fungi and bacteria (red, 287 species), or decontaminated fungi (green, 209 species). Mean performance with 99% confidence intervals (colored ribbons) and gray or lightly colored lines each denoting single repeats of ten-fold cross-validation repeated ten times. (Methods). (**B**) Per cancer type vs. healthy discrimination in the Hopkins cohort with ten-fold cross validation to calculate average performance (dots) and 95% confidence intervals (brackets). (**C**) Each cancer type vs. all others ML performance in the Hopkins plasma cohort. (**D**) Per cancer stage vs. healthy ML performance in the Hopkins plasma cohort. (**B, C, D**) Three feature sets were used: all microbial hits against the rep200 database (red; 7418 features), only fungal and bacterial species overlapping with the WIS cohort (blue; 287 species), decontaminated fungi (209 species), and for (**B, D**) an additional feature set of the top 20 ranked fungal species identified during pan-cancer vs. healthy ML (ten-fold cross-validation repeated ten times). Centered dots denote average performance and error bars represent 95% confidence intervals. Horizontal, dotted, gray or colored lines represent null AUROC and AUPR values, respectively. Null AUPR values may slightly vary between feature sets when subsetting resulted in zero-sum samples that had to be removed prior to batch correction and/or ML.

In breast cancer, we found the *Cladosporium sphaerospermum* species and the *Cladosporium* genus, previously reported in breast cancer (*62*), enriched in tumors of patients older than age 50 (Figure 13A). The *Cladosporium* genus was also found to be enriched in HER2 negative tumors (Figure 13B), although known age-HER2-status associations complicate causality (*63*). We also found the class Agaricomycetes significantly enriched in triple negative breast cancer vs. hormone receptor (HR) positive tumors (Figure 13C). Furthermore, we found significantly shorter overall survival (OS) in breast cancer patients with intratumoral *Malassezia globosa* (Figure 13D), a common fungus on human skin (*34*), in breast milk (*64*), and in pancreatic tumors, in which it was shown to have an oncogenic effect (*49*). *Malassezia restricta*, another highly abundant fungus on human skin that is also present in breast cancer, was not correlated with OS (data not shown). In lung cancer, we found higher fungal richness and enrichment of the *Aspergillus* genus and Agaricomycetes class in tumors from current smokers compared to those from never smokers (Figure 13E-F). We also found three fungal species significantly enriched in stage 3-4 over stage 1-2 lung tumors (Figure 13G). In ovarian cancer, patients with intratumoral Phaeosphaeriaceae, or its concomitant *Phaeosphaeria* genus, had significantly shorter progression free survival (PFS), shortening median PFS probability from 498 days to 135 days (Figure 13H-I). We also examined fungal associations with response to immunotherapy in metastatic melanoma. Although fungal richness did not significantly vary (p=0.88, two-sided Wilcoxon test), Capnodiales, and its genus, *Cladosporium*, were significantly enriched in non-responders (Figure 13J).

Figure 13. Fungi in the cancer mycobiome are associated with patients clinical parameters. (**A, B, C**) Differential prevalence of fungal taxa in the WIS breast tumors by age (**A**) HER2 status (**B**) or HR+ vs. TN (**C**). (**D**) Kaplan-Meier survival probability of WIS breast cancer patients positive (n=11) or negative (n=69) for *Malassezia globosa* (*P-value* from log-rank test). (**E**) Fungal richness in WIS lung tumors by smoking status. Box plot shows median line, 25^{th} and 75^{th} percentiles, and 1.5× interquartile range. (**F, G**) Differential prevalence of fungal taxa in WIS lung tumors by smoking status (**F**) or stage (**G**). (**H, I**) Kaplan-Meier plot demonstrating progression free survival (PFS) probability in WIS ovarian patients positive (n=9) or negative (n=45) for Phaeosphaeriaceae family (**H**) or positive (n=8) or negative (n=46) for the genus *Phaeosphaeria* (**I**) (*P-value* from log-rank test). (**J**) Differential prevalence of fungi in WIS melanoma tumors by response to immune checkpoint inhibitors. (**A-C, E-G, J**) *P-values* were calculated by Fisher's exact test. Only fungi that appeared in ≥5% and at least twice in one of the groups were included in the analysis. All fungi in these plots had FDR corrected values of ≤0.2.


A
Breast tumors
<=50 years (n=78)
>50 years (n=163)
p=0.02
p=0.01
Prevalence (%)
Cladosporium (g)
C. sphaerospermum (s)
B
Breast tumors
HER2− (n=181)
HER2+ (n=37)
p=0.01
Cladosporium (g)
C
Breast tumors
HR+ (n=158)
TNBC (n=22)
p=0.02
Agaricomycetes (c)
D
OS in breast cancer
Malassezia globosa Negative Positive
Survival probability (%)
p = 0.00057
Time (years)
E
Lung tumors
Never smokers (n=35)
Current smokers (n=93)
Wilcoxon
p = 0.00019
Richness
Never smokers
Current smokers
F
Lung tumors
Never smokers (n=35)
Current smokers (n=93)
p=0.04
p=0.01
Agaricomycetes (c)
Aspergillus (g)
G
Lung tumors
Stage 1−2 (n=107)
Stage 3−4 (n=18)
p=0.02
p=0.04
p=0.02
A. clavatophorus (s)
P. hubeiensis (s)
V. victoriae (s)
J
Melanoma tumors
No response (n=57)
Response (n=43)
p=0.04
p=0.04
Capnodiales (o)
Cladosporium (g)
H
PFS in ovarian cancer
Phaeosphaeriaceae Negative Positive
p = 0.0015
Recurrence probability (%)
Time (years)
I
PFS in ovarian cancer
Phaeosphaeria Negative Positive
p = 0.0019
Recurrence probability (%)
Time (years)

Intratumoral mycobiome-bacteriome-immunome interactions

Fungi commonly interact with bacteria through physical and biochemical mechanisms (*54*), motivating exploration of inter-domain co-occurrences between mycobiome and available bacteriome data in tumors (*17*, *18*). We compared presence-absence data at different taxonomic levels to shuffled counterparts, to calculate the normalized Mutual Information between domains (*65*) (Methods). Pearson correlations indicated synergistic or antagonistic interactions. Significant inter-domain co-occurrences presented in breast and lung cancers, which had the most samples, potentially reflecting a lack of power in other cancer types (Figure 14, table S8). 93% (132 of 142) of significant fungal-bacterial co-occurrences were positive, while most negative co-occurrences included the fungal family Dipodascaceae or its genus, *Yarrowia* and appeared in lung cancer (Figure 14A). In the breast cancer-specific analysis, *Aspergillus* and *Malassezia* were hubs for inter-domain co-occurrences (Figure 14B).

Fungi and bacteria were both demonstrated to elicit unique host immune responses (*24–26*, *49*, *55*, *61*, *66*, *67*), leading us to hypothesize that fungal-bacterial-immune clusters exist intratumorally. Because bacteriomes (*17*, *18*), immunomes (*68*), and mycobiomes individually demonstrate cancer type specificity, we reasoned that joined multi-species clusters also likely vary across cancer types. The TCGA data represents a unique opportunity to compare between the mycobiome, microbiome and host information, and so we could test these hypotheses. We compared fungal and bacterial genera overlapping from the WIS cohort in the TCGA cohort with concomitant TCGA immune cell compositions derived from CIBERSORT (*68*, *69*), using MMvec (microbe–metabolite vectors), a neural network architecture previously developed to estimate microbiome-metabolite co-occurrences (Methods) (*70*).

Unsupervised analyses revealed three distinct fungi-bacteria-immune clusters driven by fungal co-occurrences, herein named F1 (*Malassezia-Ramularia-Trichosporon*), F2 (*Aspergillus-Candida*), and F3 (multi-genera including *Yarrowia*) "mycotypes" (Figure 15) (Methods). These fungal mycotypes show distinct patterns of co-occurrence with the bacteria and immune cells within the tumors. F1 and F2 mycotypes comprised fewer but more prevalent fungal genera in TCGA. These mycotypes show similar clustering of immune cells but differentiated in the bacterial interactions. F3 represents a larger number of fungi but with lower prevalence compared to F1 and F2 mycotypes and are

associated with higher levels of innate immune cells. Further analysis into these interactions is necessary.

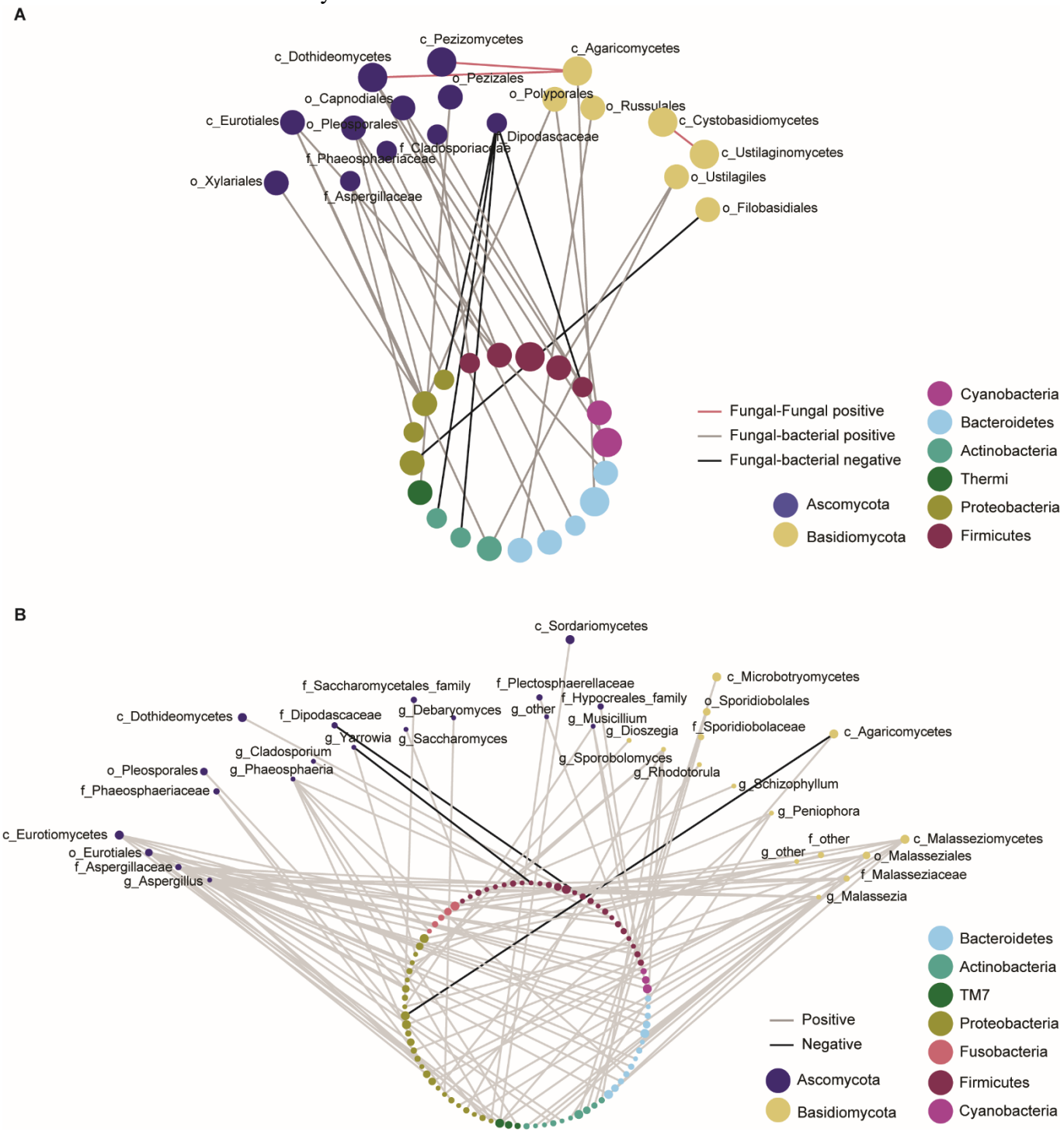


Figure 14. Establishing cancer specific mycobiome-bacteriome interactions. (**A, B**) Networks of fungi-bacteria co-occurrences at different taxonomic levels found in WIS (**A**) lung tumors or (**B**) breast tumors drawn with Cytoscape (3.8.1). Nodes represent taxa color coded according to phyla. Bacterial nodes are organized by phyla in a circle at the bottom. Fungal nodes are labeled with taxonomy preceded by a letter representing the taxonomic level (g: genus, f: family, o: order, c: class) and are organized in rainbow fashion according to taxonomic level from class (outer layer) to genus (inner). Nodes are grouped on a gradient from outer to inner layer based on taxonomic hierarchy. The size of the node corresponds to its taxonomic level from class (large) to genus (small). Edges represent co-occurrences between nodes with significant Normalized Mutual Information scores (permutation test n=1000, BH-FDR<0.25) (Methods). Color represents fungal- bacterial positive (gray) and negative (black) interactions or positive fungal-fungal (pink) interactions.

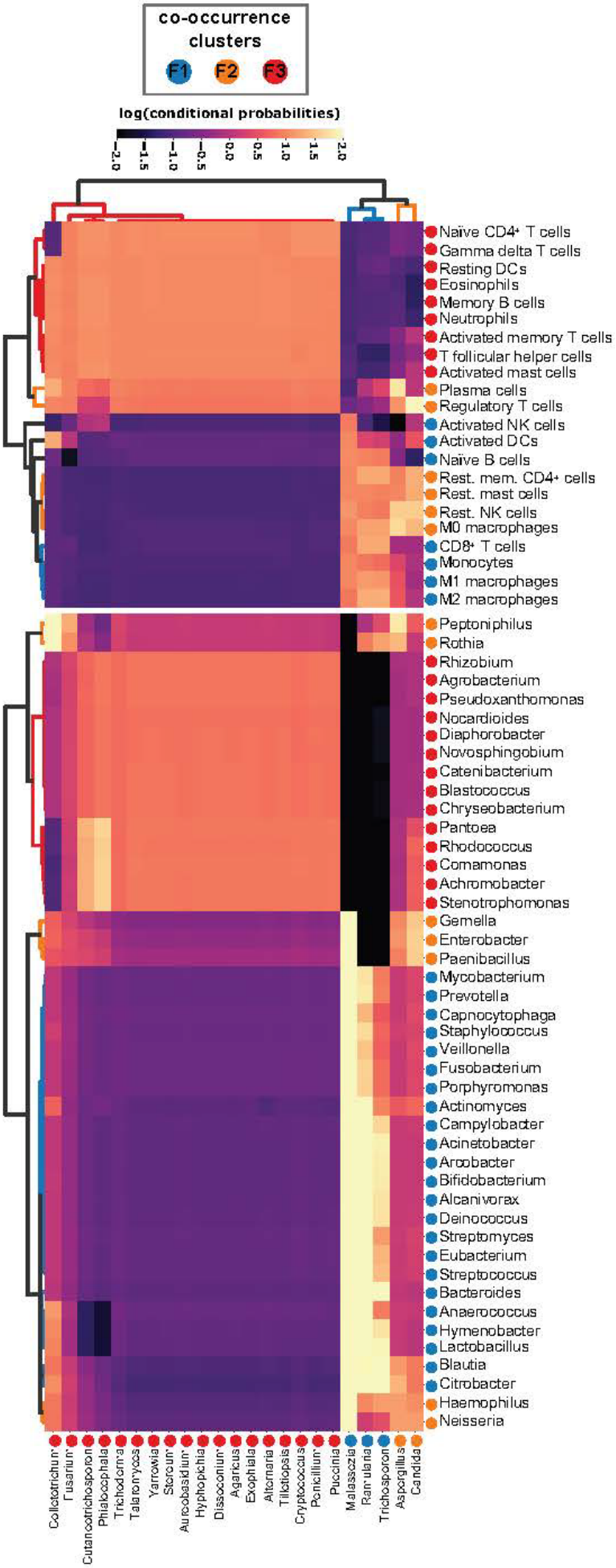


Figure 15. Establishing pan-cancer mycotypes through mycobiome-bacteriome-immunome interactions. Co-occurrence analyses of TCGA fungal and immune cell compositions (*40*), and bacterial abundances at the genus level using MMvec (*42*). Only WIS-overlapping fungal and bacterial genera were included (see Methods). Hierarchical clustering linkage information identified three distinct clusters ("mycotypes") associated with groups of fungal genera: F1, F2, and F3.

ITS2 sequencing optimization and validation in the WIS cohort

Our approach to detecting and characterizing intratumoral fungi was by amplification and sequencing of the internal transcribed spacer 2 (ITS2). The ITS2 is a region within the fungal rDNA cistron. While this region is transcribed, it is not active within the ribosome and is thus much less conserved than the active rDNA sequences (18S, 5.8S, 28S). Hence, it can be used as a fungal species phylogenetic "barcode" (*71*), akin to the 16S sequence in bacteria. However, there are several issues accompanying the use of this region as a fungal barcode. First, the region varies in length between different fungi, mainly ranging between 200-500 bases (*72*). This difference in size might cause a bias against species with a long ITS2 region, mainly during the sequencing step (*73*). In addition, the rDNA copy number within fungi can vary greatly, with species reported to have one copy and others reported to have 200 copies within a single genome (*72*). Fully accounting for this potential bias is difficult as the copy number for most species is unknown, and can differ even among strains of the same species and during the life cycle of a single strain (*74*). Hence, many of the analyses we performed were done with presence/absence data to avoid the effect of such bias.

ITS2 sequencing pipeline was extensively optimized and validated prior to WIS tumor cohort analysis. This was done by carefully choosing the ITS2 primers by *in-silico*, *in-vitro* and literature based evidence; strict decontamination with negative control samples; flooring of reads to increase signal to noise ratio; sequencing validation with fungal mock community controls; and triplicate technical repeats of tumor samples to test for reproducibility, as described below.

ITS2 fungal primers capture most of the fungal kingdom

The primers used were chosen out of the available primers in the literature after *in-silico* and *in-vitro* analysis of several primer pairs (Figure 16). *In-silico*, these primers captured (with up to one mismatch) on average (between different databases) ~80% of the fungal kingdom (Table 3). In addition, they captured only 3.3% of the outgroup database, which included mainly plants, indicating high specificity for the fungal kingdom. Furthermore, PCR testing of these primers on several fungal species spanning the two major fungi phyla (Ascomycota and Basidiomycota) gave the best results out of the primer pairs tested (Figure 16C). Finally, a study by Beeck et al. (2014) demonstrated that these primers are superior to the more commonly used primers in

amplifying fungi in soil samples (*75*), strengthening our results. In our tumor cohort, 70% of the reads were classified to species level with the ITS2 sequencing and classification pipeline (Table 4). This is a great improvement on past classification pipelines in which as low as 2.6% of total reads were classified to species level (*76*).

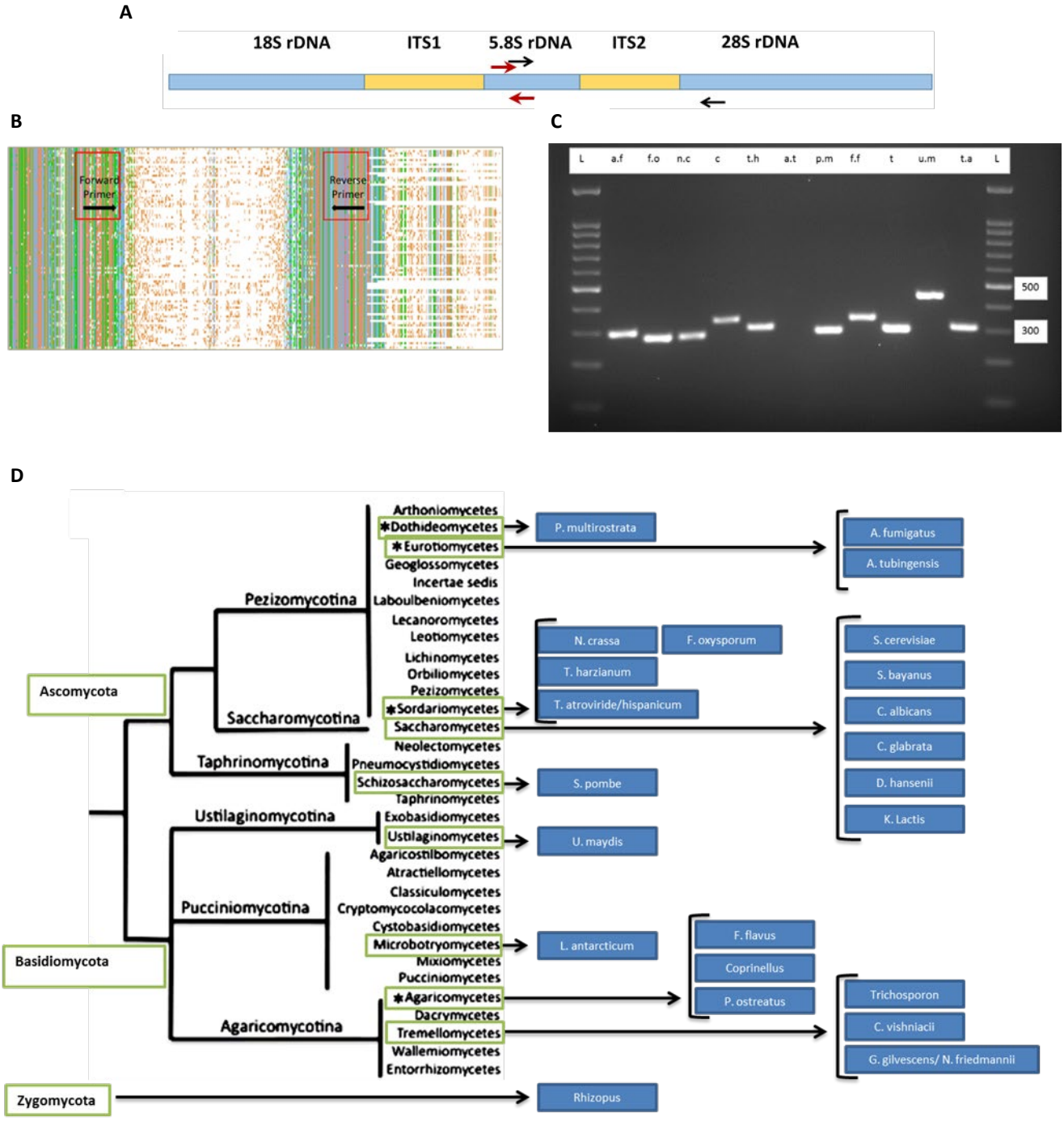


Figure 4. ITS86F-ITS4 primers were tested *in-silico* and *in-vitro*. (A) The fungal rDNA region. ITS86F – ITS4 primers (black) used for sequencing. ITS3 – ITS86R primers (red) used for qPCR. **(B)** Multiple alignment of the fungal rDNA from many fungal species. Primers are anchored in conserved regions and the amplicon spans the less conserved ITS2. **(C)** The fungi tested span the two main phyla in the fungal kingdom. **(D)** Fungal phylogenetic tree with fungi tested for ITS2 amplification by the primers depicted in blue squares. *a.f - Aspergillus fumigatus; f.o - Fusarium oxysporum; n.c - Neurospora crassa; c - Coprinellus sp.; t.h - Trichoderma harzianum; a.t - Aspergillus tubingensis; p.m - Phoma multirostrata; f.f - Flavodon flavus; t - Trichosporon sp.; u.m - Ustilago maydis; t.a - Trichoderma atroviride*

Primers	ITS86F-ITS4R				
Mismatches		0	1	2	3
Fungal databases	ISHAM	52.82	75.42	79.24	86.72
	THF	61.45	87.08	89.43	93.74
	FINDLEY	59.67	72.07	75.32	78.67
	RTL	61.38	92.63	95.58	96.7
	SILVA_LSU	69.21	84.94	87.7	91.42
	SILVA_SSU	80	87	96	98
	UNITE	66.15	79.18	82.75	85.78
Mean		64.38	82.62	86.57	90.15
Outgroup database	SILVA_LSU	0.44	3.3	46.67	61.64

Table 3. Percentage of sequences matching the sequencing primers in different fungal databases

Taxonomic rank	# Reads	Percent
Kingdom	59,919,467	100
Phylum	59,412,527	99.15
Class	59,200,179	98.8
Order	59,035,172	98.52
Family	54,747,190	91.37
Genus	53,154,728	88.71
Species	42,155,685	70.35

Table 4. Number and percent of the fungal reads that were classified to each taxonomic level. All fungal reads were included in this analysis before flooring and normalization.

Negative control samples enabled data clean up and decontamination

When working with low biomass samples, negative controls are of great importance. In low biomass samples, the contamination may be stronger than the signal itself. Hence, negative controls allow segregation of the signal from the noise. Our cohort includes two types of negative controls (table S5): (1) 191 DNA extraction controls performed on empty tubes (with DDW only) in parallel to sample DNA extraction (negative controls for contaminations introduced during lab processing) and (2) 104 paraffin controls which were made by sampling paraffin only (without tissue) from a subset of the study paraffin blocks of tumor tissue (negative controls for center contaminations). The 295 negative controls allowed for better understanding of the fungal signal in the tissues vs. background noise as can be detected in the negative control samples. The

histogram of the number of reads per amplicon sequence variant (ASV) per sample as well as the number of reads per sample (Figure 17A-B) both presented a bimodal distribution with the peaks found on either side of 1000 reads/ASV or 1000 reads/sample. We found that the chance of an ASV to have more than 1000 reads was 3 times higher in samples vs controls (21.6% vs 7.1%). We therefore, used such a delineation between signal and contamination noise and floored the data such that any ASV per sample with <1000 reads was converted to 0 reads.

The negative control samples were then used to flag potential contaminant species (Figure 17C). Out of 456 species detected from 1191 ASV's in the data, 13 species unique to the negative control samples were removed from the dataset (Figure 17D). For an additional 63 species that were detected in both negative control samples and true samples, statistical testing was applied in two distinct steps using the extraction control samples and the paraffin control samples separately (Methods). Forty-two species (out of the 63 that were tested) passed both filtering steps in at least one condition. All of these 42 species, as well as the 380 species that did not appear in any of the 295 controls were considered part of the 'fungal world' that was used for all downstream analysis of the WIS cohort. The same filtering steps were also performed for each of the taxonomic levels (table S4). Note that only fungi and bacteria that passed the filtering steps in at least one of the tumor types were included in most of the analysis in this work. In total 4.6% (21/456 species) of fungal species detected in our samples were determined to be potential contaminations, this is as opposed to 94.3% of bacterial species in the same samples (*17*). This suggests the mycobiome is less prone to contaminations relative to the microbiome analysis in low biomass samples such as tumors.

Fungal ITS2 sequencing successfully captures fungi in Mock communities

A mock community of 17 fungal species was generated to validate the ITS2 experimental procedure and assess the success of detecting different fungi. Fourteen out of the 17 species were detected (Table 5). One of the species that was not detected (*Flavodon flavus*) was wrongly classified to a different family in the same order (Polyporales). Overall, 99.89% of the reads belonged to the species included in the mock. We repeated the ITS2 amplification and sequencing two more times and reached almost identical results, detecting the same fungal species (data not shown).

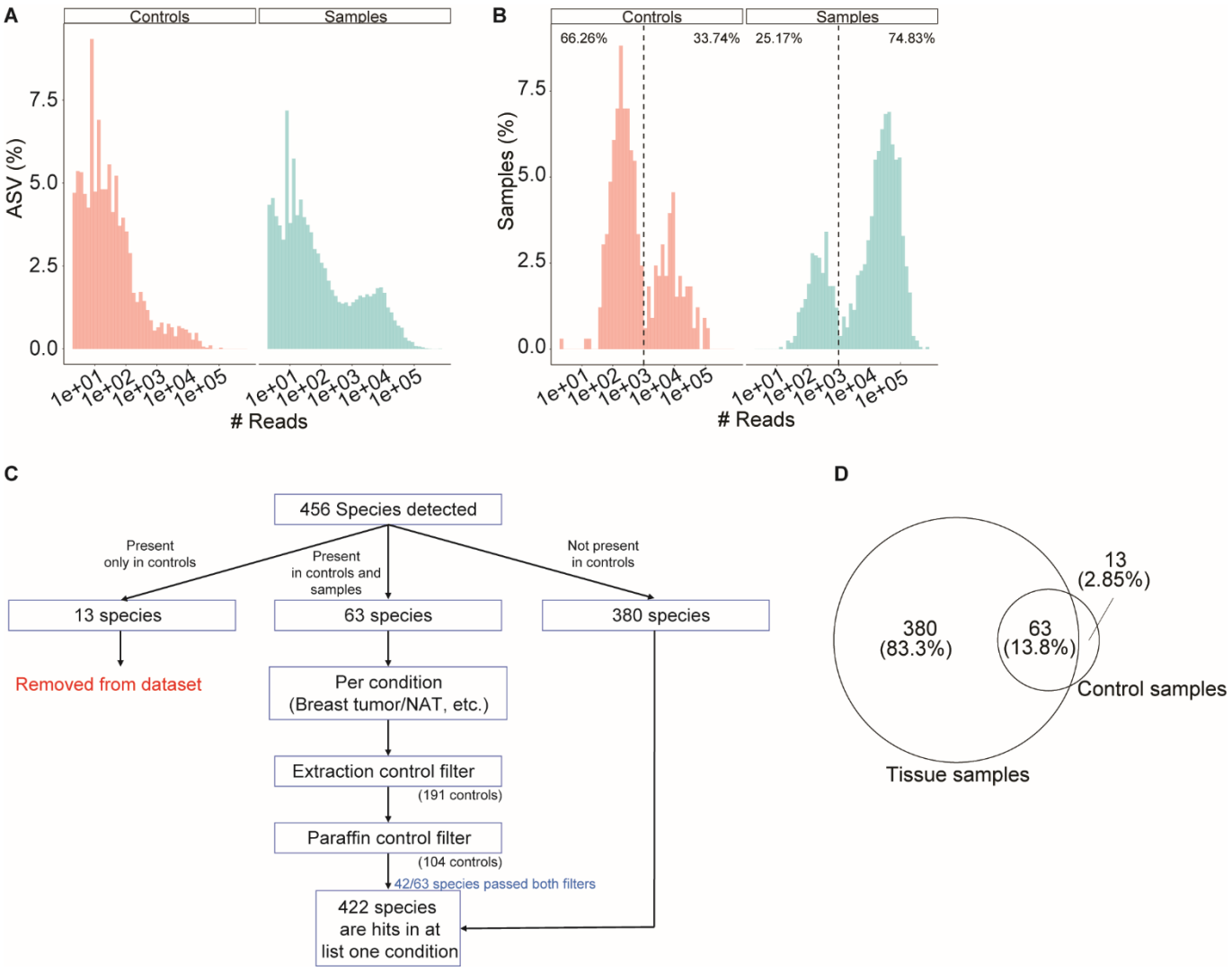


Figure 17. ITS2 sequencing pipeline development. (**A**) Histograms of the number of reads per ASV per sample in control samples (extraction controls and paraffin controls) vs. all other samples. (**B**) Histograms of the total number of fungal reads before flooring and normalization per sample in control samples (extraction controls and paraffin controls) vs. all other samples. (**C**) Schematic illustration of the decontamination workflow applied to ITS2 data to flag and remove contaminant species. (**D**) Venn diagram of the overlap between all species before hit calling, as detected in control samples and tissue samples.

	Species spiked	Detected Yes/No	Taxa level classified
1	*A. laibachii*	Yes	Species
2	*C. albicans*	Yes	Species
3	*C. glabrata*	Yes	Species
4	*C. micaceus*	Yes	Species
5	*G. antarctica*	Yes	Species
6	*K. lactis*	Yes	Species
7	*P. prolifica*	Yes	Species
8	*R. arrhizus*	Yes	Species
9	*S. cerevisiae*	Yes	Species
10	*S. pombe*	Yes	Species
11	*A. fumigatus*	Yes	Genus
12	*A. tubingensis*	Yes	Genus
13	*N. crassa*	Yes	Family
14	*P. multirostrata*	Yes	Family
15	*D. hansenii*	No	N/D
16	*F. flavus*	No	N/D (*)
17	*S. bayanus*	No	N/D

Table 5. Fungal species in mock samples that were detected by the ITS2 sequencing pipeline. 1.9×10^{-5} ng of DNA from each of 17 fungal species were pooled together and spiked into 100 ng of human DNA. Detection status and taxonomy level of classification are depicted in the table. *This species was wrongly classified by our pipeline as Polyporales (o) Irpicaceae (f) unknown genus.

Fungal ITS2 sequencing in tumor samples is reproducible

To assess the reproducibility of our technical and computational pipeline we repeated the ITS2 amplification and sequencing three times, for 88 human tumor or NAT samples. The number of reads and number of ASV's received per repeat significantly correlated between repeats (see Figure 18A-B for representative figures). For 82 samples that passed quality control, we compared the Bray-Curtis dissimilarity scores between all pairs that belong to the same original sample versus all pairs that belong to different samples within the same tissue type. We found that the dissimilarity was significantly lower between repeats relative to between samples from the same tissue (p-value$<2.22\times10^{-16}$) (Figure 18C). The similarity grew higher (dissimilarity lower) when the sample repeats had higher read counts (Figure 18D) but did not differ based on the number of ASV's in the samples (Figure 18E).

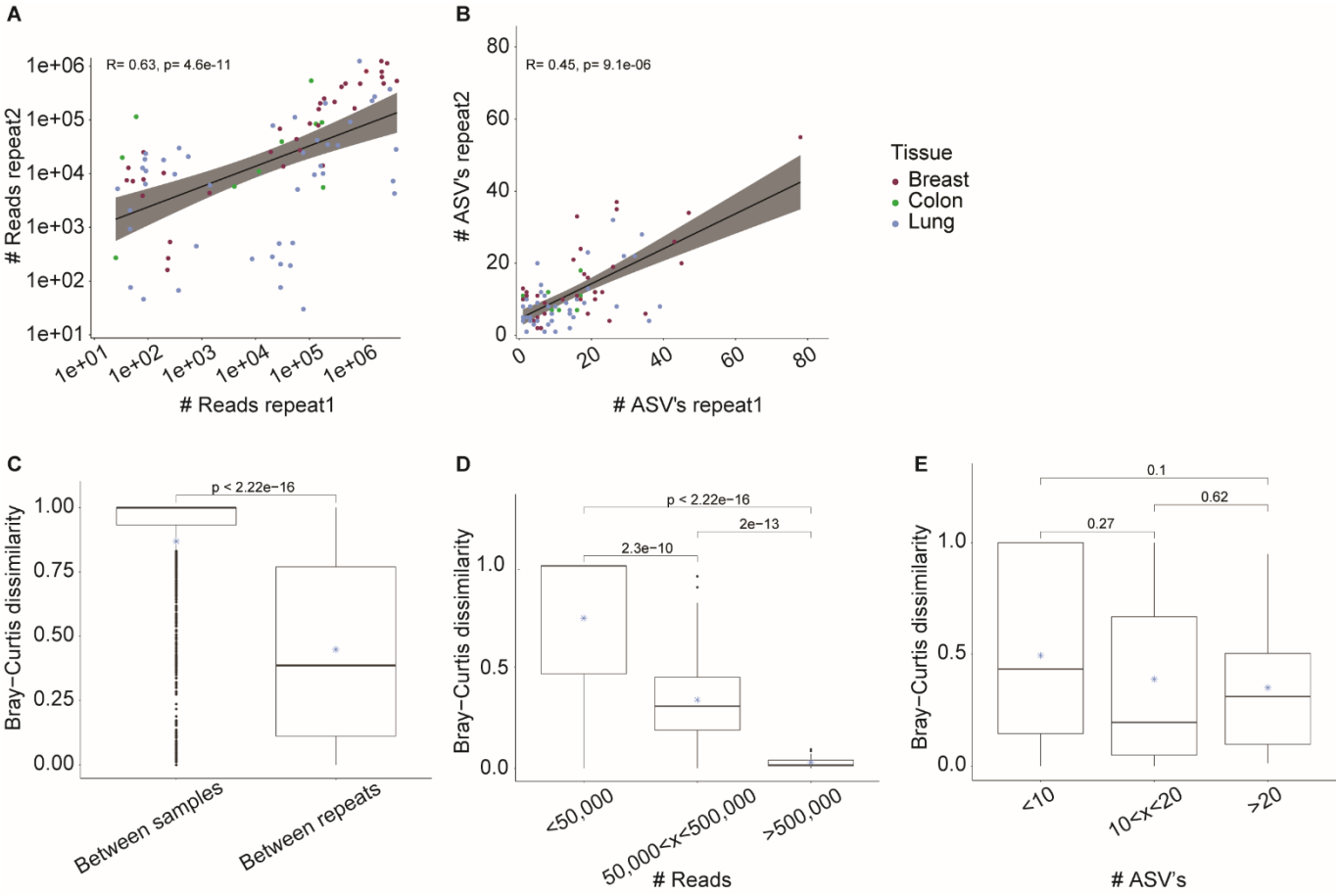


Figure 18. Fungal ITS2 sequencing in tumor samples is reproducible (**A, B**) Scatter plot demonstrating the Pearson correlation between the number of total ITS2 fungal reads sequenced (**A**) or the number of ASV's detected (**B**) in two technical repeats of sample sequencing. Regression lines and confidence intervals are shown. Pearson correlation coefficient (R) and *P-value* (p) are presented. (**C**) Box plot of Bray-Curtis dissimilarity scores comparing the dissimilarity between the fungal composition in pairs of samples (both tumor and NAT) from the same tissue type vs. the dissimilarity between the fungal compositions in sequencing technical repeats of the same samples. Samples represent breast (tumors, n=17; NAT, n=16), colon (tumors, n=4; NAT, n=5) and lung (tumors, n=20; NAT, n=20) tissues. DNA from each sample was amplified and sequenced in triplicates. (**D, E**) Box plots of Bray-Curtis dissimilarity scores comparing the dissimilarity between the fungal composition in sequencing technical repeats of the same samples divided by the number of reads of the sample (**D**) or the number of ASV's in the sample (**E**) . (**C-E**) Two-sided T-test was performed; *P-value*s are depicted in the plots. The asterisks depict the mean. Box plots shows median line, 25th and 75th percentiles, and 1.5× interquartile range.

Functional analysis of all microorganisms in tumors

The above genomics approach focuses on the taxonomy of the microorganisms within the tissue. DNA sequencing (shotgun and PCR based) enables the detection and identification of the microorganisms that are present within our samples. There is accumulating evidence that as the identity of the organisms within different samples may greatly differ, as seen in our samples as well (Figure 7D,E), the functional patterns or capabilities of these organisms may be more coherent within different test groups (*1*). Where possible, shotgun sequencing provides additional information on the genes present within the sample. However, this is not functional information, since it does not verify which of the genes detected is expressed. In contrast, RNA sequencing provides the functional information for the entire sample, hence sequencing of RNA in such metagenomics samples could be very informative. However this approach is not always possible due to the higher abundance of host (human) RNA relative to the RNA of microorganisms in the tissue. In the TCGA cohort, while many of the samples (>10,000) come from RNA sequencing experiments, they suffer from a very low number of reads per sample (mean ~150). Hence, functional analysis was not possible with this data, affirming the need for a different solution for microorganism functional analysis in low biomass samples, such as tumor samples. There have been tools developed to predict the functional capacity of the bacteriome based on 16S sequencing such as PICRUSt (*77*). This method relies on comprehensive knowledge of the bacterial genomes, knowledge that at the moment, is lacking for most fungi.

To further our understanding of the function of the fungi and all microorganisms within the tumor samples, I am developing a protocol for human RNA depletion from RNA samples. The depletion of the human RNA will allow for the detection and sequencing of all non-human RNA within low biomass samples, such as human tumors. This will allow for a comprehensive viewpoint of the functions of the microorganisms within the tumor. Together with the identification of the fungi by ITS2 sequencing, and the bacterial identification (*17*), we will have a better understanding of not only the identity, but also the functions of microorganisms within tumors.

In order to deplete the human RNA before RNA–seq, we optimized the following method (Figure 19A): First, total RNA is generated from a human cell line that was validated to contain no microorganisms (e.g. mycoplasma) and is fragmented into short

oligonucleotides. cDNA is then synthesized from this RNA and hybridized to total RNA that was prepared from a tumor sample. After an annealing step, Ribonuclease H (RNase H) - a non-sequence-specific endonuclease enzyme that catalyzes the cleavage of RNA in RNA/DNA substrates - is added, to cleave any RNA that hybridized with a cDNA. As no RNA from microorganisms is found in the cell line, the RNase H preferentially degrades the human RNA, leaving all non-human RNA intact. At this stage, DNase 1 is added to degrade the cDNA, leaving behind only non-human RNA. This RNA can then be sequenced using standard protocols and the functions of the microorganisms within the tumor can then be explored.

In a preliminary attempt, cDNA probes were prepared from the HS5 human cell line. Human RNA was then depleted from a sample containing a mix of the human HS5 cell line, the fungal *S. cerevisiae* and the bacterial *E.coli* RNAs that were mixed at a ratio which represents 100:1:1 cells (based on RNA content per cell from bionumbers: http://bionumbers.hms.harvard.edu/). Two probe concentrations and a no probe negative control were used. Depletion efficiency was tested by qPCR of four genes – human 18S, *S. cerevisiae* 18S, *E. coli* 16S and human PABPC1. For the first three genes, an additional control was performed to estimate the efficiency of DNA degradation during RNA extraction. For this purpose, qPCR was performed directly on the RNA (before cDNA synthesis step).

Based on the human 18S qPCR performed on the RNA samples, we found that the degradation of the cDNA probes was not successful in our preliminary experiments (Figure 19B). The remaining cDNA probes thus affected the levels of human 18S detected in the cDNA samples, and skewed the estimation of human RNA depletion. The majority of the human 18S signal seems to originate from the non-degraded DNA probes and not from human RNA left un-degraded in the sample of interest. Taking this into consideration it seems like the depletion of the human 18S gene was successful without any effect on yeast and bacterial RNA. The PABPC1 results are not clear, since the RNA qPCR control was not performed. We will further calibrate the method to optimize probe degradation. The method will then be used on RNA from human tumors to characterize the functions of the microorganisms within the samples. These experiments will complement the data from 16S and ITS2 sequencing characterizing the bacterial and fungal species within the samples.

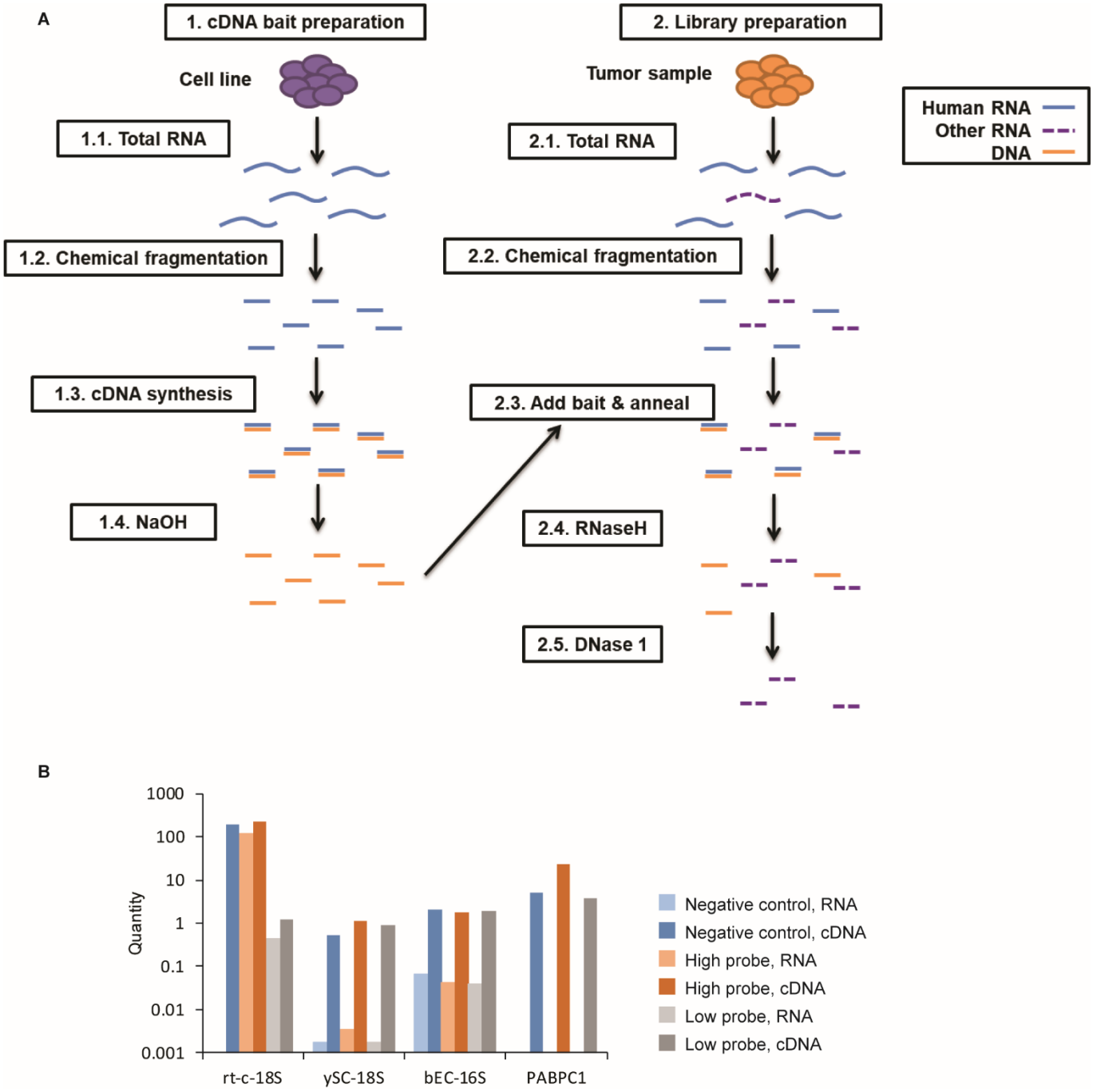


Figure 19. Human RNA depletion. (**A**) Outline of the human depletion protocol. Human RNA in tumor samples is depleted by cDNA probes. These probes are prepared from human cell line total RNA. The specific degradation of human RNA elevates the concentration of microorganism RNA. Following this protocol, the microorganism RNA can be sequenced and the functions of the microorganisms within the tumor can be explored. (**B**) qPCR results from preliminary human depletion experiment. The degradation of the cDNA probes was not successful. The cDNA probes remaining affected the levels of human 18S detected in the cDNA samples, and skewed the estimation of human RNA depletion. The majority of the human 18S signal seems to originate from the non-degraded DNA probes and not from human RNA left un-degraded in the sample of interest. Taking this into consideration it seems like the depletion of human 18S gene was successful without any effect on yeast and bacterial RNA.

Discussion

Here we strived to create a pan-cancer atlas of the fungal composition across multiple human cancer types. We characterized the mycobiomes of 17,401 tissue and blood samples in four independent cohorts and 35 cancer types with complementary strategies. Fungal DNA (qPCR, ITS2 sequencing, TCGA WGS), RNA (FISH, TCGA sequencing data), polysaccharides and proteins (Stainings) were detected in human tumors. Fungi were detected mainly intracellularly in either cancer or immune cells. In addition, our different data modalities revealed ubiquitous, cancer type-specific fungal ecologies with lower diversities and abundances than matched bacteriomes. The cancer-type specific fungal profiles were similar to the profiles detected in the NAT of the same corresponding tissue or organ. Furthermore, we found that tumor and NAT from the same patient were more similar than samples from different patients with the same cancer type, suggesting within patient specific mycobiome profiles. These profiles were used to discriminate between cancer patients based on tumor type as well as between cancer patients and healthy individuals even when using fungal profiles from the blood. We found significant association between fungi and clinical patient data suggesting fungi may be used as clinical biomarkers for breast subtype, response to therapy and prognosis. Finally, we also detected significant co-occurrences between fungal and bacterial taxa in the different tumor types. This analysis provides a comprehensive characterization of the tumor mycobiome in many tumor types and leads to many open questions that now remain to be answered with further research.

Our study has several technical caveats. While intratumoral fungal presence was demonstrated by four different staining methods, and revealed tumor-specific localization patterns, fungal staining in tissues proved challenging, with sensitivity and specificity of the four distinct methods varying across cancer types. Further work is necessary to elucidate fully the fungal localization within the tumors. In addition, our analysis does not inform whether the detected fungal components in the tumors correspond to live or dead fungi. With the exception of a few rare cases where fungal cells were visualized, the majority of fungal staining detected within the tumors was intracellular and lacked fungal cell morphology. There is evidence of fungi living within human macrophages after phagocytosis (*78*, *79*). Whether the fungal stains we detect here are due to live fungal internalization and degradation by the human cells

remains to be explored. Furthermore, thus far we only rarely cultivated fungi from tumor tissue. We are currently optimizing our protocols for fungal tissue cultivation.

We focused on ITS2 sequencing to determine the fungal species within our tumor samples. However, recent debate within the mycology community suggests that different fungal barcodes result in detection of different fungal profiles within samples (*80*). Furthermore, the variance in ITS2 length and copy number per fungal species (*72*, *73*) can bias our estimation of the relative abundance of fungal content. In particular, the abundance of fungi with more copies of the ITS region may be overestimated, while fungi with very long ITS2 regions may be underestimated since their ITS region may be less efficiently PCR amplified and sequenced. To overcome these difficulties we intend to re-sequence our samples using additional fungal barcodes such as the ITS1 and 18S to get a more balanced and reliable picture of the fungal profiles within the tumors.

In our ITS2 sequencing, we detected a bimodal distribution of the number of reads per sample, with 66.26% of control samples having under 1000 reads and 74.83% of true samples having over 1000 reads. A close look at the reads per ASV revealed that this bimodality is mainly due to index-hoping during sequencing. To test this hypothesis and overcome it, we can use dual indexes per sample in future experiments. This should potentially eliminate the index-hoping and remove the need for the significant flooring of read counts that we performed in this analysis. Finally, fungal amplicon sequencing proved less prone to contaminations relative to bacterial sequencing. Only 4.6% of fungal species were removed as potential contaminations, relative to 94.3% of bacterial species in the same samples (*17*). The difference between the levels of contaminants may stem from several reasons. Fungi are estimated to have at least two orders of magnitude less cells per cubic meter of air relative to bacterial cells (*81*). In addition, many enzymes used during experimental procedures are derived from bacterial cells, giving rise to specific bacterial contaminations from laboratory materials that are used during sample processing.

In addition to the ITS2 sequencing in the WIS cohort, our study also used data from existing cohorts, mainly the TCGA. The different cohorts used in this study strengthen the results described above. However, distinct differences in the data also emerge. The richness within samples is at least an order of magnitude higher in the TCGA data

relative to the WIS data. This is most likely due to careful curation of the WIS cohort using negative controls as well as to read splitting during shotgun metagenomic alignments in the TCGA cohort, which potentially inflates the number of species detected per sample. In addition, the number of overlapping species between the cohorts was small, likely due to the difference in fungal content in the databases used for classification. Despite this, tumor-derived species from our cohort (WIS) provided nearly equivalent discriminatory performance as a multi-domain database 26-fold larger, suggesting a significant tumor origin, generalizable across additional cohorts from different continents. The approach used on the TCGA data has a great advantage in that comparisons to human data within the samples are possible. In addition, this method could be used on any available WGS/RNA-seq dataset. However, the datasets used must have sequenced each sample deep enough to receive a significant number of fungal reads that will enable proper analysis. Furthermore, these datasets do not include negative controls and so decontaminations can only be performed by bioinformatic and statistical methods. Finally, sample-processing starting from sample collection to nucleic acid extraction were not performed with microorganisms in mind and so these datasets are more prone to contaminations and likely give a partial picture of the sample mycobiome due to non-ideal nucleic acids extraction methods and low read count per sample.

Many interesting questions arise from this analysis and demand further research. The cancer-type specific fungal profiles were similar to the profiles detected in NAT, raising the question of the source of the tumor mycobiome. The lack of enough true normal tissues in our cohorts limited our ability to determine if the source of intratumoral fungi is their surrounding normal tissue or if fungi found in tumor-adjacent normal tissue originate from the tumor fungal community. However, the detection of specific fungal species in tumors that are known to exist in the normal tissue of the same organ as well as the clustering of breast normal samples with breast NAT and tumor samples in our data, point towards the source of the tumor mycobiome being, at least in part, in the normal resident tissue. Further research is needed to fully elucidate the source of the tumor mycobiome in the different tumor types.

While broadening the cancer microbiome landscape, our findings do not inform if there are fungi that causally impact or complicitly aid carcinogenesis. The fungi detected may have an active role and effect on the tumor, or they may be hitchhikers of the lower

immune surveillance in most tumors. In both cases, they may be used as biomarkers for the cancerous state. While evidence of the tumor promoting or inhibiting functions of fungi or fungal components already exists (*49*, *50*, *82*), further research is necessary to better understand their effects in different tumor types. *Malassezia globosa*, which we found to be significantly enriched in breast tumors with worse overall survival, was previously shown to have oncogenic effects in PDAC in mice (*49*). In addition, higher levels of *Malassezia* were detected in colorectal cancer vs. healthy individuals (*42*). Whether it has a similar effect in breast cancer demands further research. Another interesting finding was the enrichment of *Cladosporium* in melanoma tumors from patients that did not respond to ICI. It is important to validate these results using another non-related cohort as well as to find if this fungus plays a causal role in the response of a patient to the treatment.

We also provide an analysis of plasma microbiomes in treatment-naive, early-stage cancers, suggesting the use of microbial nucleic acids in the plasma as an early biomarker for cancer detection. Microbial profiles from cell free DNA in patients' blood could potentially be used to diagnose cancer and cancer type in individuals. The robustness of microbial-augmented liquid biopsies in other cohorts and among diseased, non-cancer-bearing or infected hosts remains to be characterized.

The tumor can be regarded as an ecological niche. Strong positive correlations between fungal and bacterial richness, abundances, and co-occurrences across many cancer types portray the tumor microenvironment (TME) as a non-competitive space for microbial colonization. This situation differs from the gut, where competitive interactions often dominate, especially under anti-cancer or antibiotic therapies (*55*, *83*). It remains unclear whether this permissive phenotype is passively allowed by immunosuppressed TMEs (*84*), or represents an active pursuit for greater ecosystem multifunctionality (*85*) or selection advantages for tumors (*8*, *9*, *49*). We revealed many significant fungal-bacterial co-occurrences, many of which merit additional study to understand fully the fungal-bacterial relationship.

Finally, whereas our study comprehensively characterizes the fungal species within tumor tissue we are mostly blind to the functions these fungi perform in the tumor setting. As we detected high fungal diversity between tumor samples from the same tissue it is of great interest to test whether the functions they perform may show higher

levels of similarity. To this end, I am developing a protocol for host RNA depletion that will enable us to focus on the functions the microorganisms perform within the tumor.

The tumor mycobiome is a new and exciting field in tumor microbiome research. This first pan-cancer mycobiome atlas informs future directions of study while characterizing a new layer of information in cancer. The full potential of the tumor mycobiome as a diagnostic and prognostic tool as well as the effects it may have on tumor progression and response to treatment demand further endeavors.

Materials and methods

Weizmann cohort

Sample collection

The samples of the ITS2 cohort were collected from nine medical centers, and their DNA extraction as well bacterial characterization were reported by Nejman et al, 2020 (*17*). For ITS2 profiling, 1183 samples of this original cohort were used (Table 1; table S5). Samples include tumor, normal adjacent tissue (NAT) and normal tissue from eight tumor types for a total of 12 conditions (condition is defined by the tissue type and the tumor/NAT/normal status) (Table 1). Samples included both formalin fixed paraffin embedded (FFPE) and snap frozen samples. To account for potential contamination by fungi or fungal DNA during sample handling and processing, our cohort also included two types of negative controls: 104 paraffin-only controls which were made by sampling paraffin only (without tissue) from the study FFPE blocks and 191 DNA-extraction negative controls in which only sterile DDW was introduced at the beginning of the DNA-extraction pipeline. These controls enabled detection of potential fungal contaminants and delineation of signal versus noise allowing for appropriate processing of the data prior to analysis (see below). Note that matching bacterial data of the same samples that was used in this study, was generated by us and published in Nejman et al. (*17*).

ITS2 amplification and sequencing

ITS2 sequencing was used for fungal identification in all samples. PCR was performed on 100ng of DNA per sample (or the maximum available). For extraction controls a

volume of 5ul per sample was used, and for empty paraffin controls a volume equal to the volume taken for the matching sample was used. Forward primer ITS86F 5'-GTGAATCATCGAATCTTTGAA-3' and reverse primer ITS4 with rd2 Illumina adaptor 5'-AGACGTGTGCTCTTCCGATCT - TCCTCCGCTTATTGATATGC-3' were used for the first PCR amplification. PCR mix per sample contained 5ul sample DNA, 0.2uM per primer (primers purchased from sigma), 0.02unit/ul of Phusion Hot Start II DNA Polymerase (Thermo Scientific F549), 10ul of X5 Phusion HS HF buffer, 0.2mM dNTPs (Larova GmbH), 31.5ul ultra pure water, for a total reaction volume of 50ul. PCR conditions used were 98°C 2min, (98°C 10 sec, 55°C 15 sec, 72°C 35sec) X 35, 72°C 5 min. A second PCR was performed to attach Illumina adaptors and barcode per sample for 6 additional cycles. Samples from 1st PCR were diluted 10 fold and added to the PCR mix as described above. Primers of second PCR included: forward primer P5-rd1-ITS86F 5' - AATGATACGGCGACCACCGAGATCT - ACACTCTTTCCCTACACGACGCTCTTCCGATCT - GTGAATCATCGAATCTTTGAA-3', and reverse primer 5'-CAAGCAGAAGACGGCATACGAGAT - NNNNNNNN - GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'. Every 96 samples were combined for a single Mix by adding 14ul from each. Before mixing, an aliquot from each of the samples was run on an agarose gel. In cases where the amplified bands were very strong, samples were diluted between 5 and 20-fold before they were added to the mix (table S3). Each sample mix was cleaned with Qiaquick PCR purification kit (Qiagen, catalog number 28104). Four cleaned sample mixes were then combined into a single mix of 384 samples, and size selection was performed with Agencourt AMPure XP beads (Beckman Coulter #A63881) to remove any excess primers. Beads to sample ratio was 0.85 to 1. Samples were then run on the Miseq v3 600 cycles paired end with 30% PhiX. Overall, six runs of Miseq were performed for this study.

ITS2 sequencing analysis

ITS2 read classification pipeline

The ITS2 classification pipeline was built with python 3.6. For each sequencing library, paired end reads were joined using PEAR (version 0.9.10) followed by filtering of merged reads by minimum length of 80bp and trimming of primers from both ends with cutadapt (version 1.17). Within the QIIME 2 environment (version 2018.8), Dada2 was

used to create amplicon sequence variants (ASV's), then ITSx (version 1.1b1) was used to delineate ASVs to ITS2 regions (removing preceding 5.8S and trailing 28S sequences). A taxonomic naive bayesian classifier was trained on the UNITE database (version 8, dynamic, sh_taxonomy_qiime_ver8_dynamic_04.02.2020.txt) and used to classify the processed ASVs. ASVs were filtered by the ITSx and UNITE classifications to include fungal reads only. Any ASVs that were classified by ITSx as fungi were included in the downstream analysis. Any ASVs that were classified by ITSx as non-fungal, were included in the downstream analysis only if their classification as fungi reached a class or lower phylogenetic level by UNITE. Seventy percent of ASV reads that were included in the downstream analysis were classified to species level (Table 4). The other 30% of ASV reads were classified to higher taxonomic levels.

ITS2 data flooring and normalization

The histogram of the number of reads per ASV per sample as well as the number of reads per sample (Figure 17A-B) both presented a bimodal distribution with the peaks found on either side of 1000 reads/ASV or 1000 reads/sample. We found that the chance of an ASV to have more than 1000 reads was 3 times higher in samples vs controls (21.6% vs 7.1%)We therefore floored the data such as any ASV per sample with <1000 reads was converted to 0 reads.Next, we introduced two types of data normalization: (1) Library normalization: samples were normalized to account for the difference in the average number of reads/sample per library. A factor was assigned to each of the six sequencing libraries to reflect the fold change of the mean number of reads/sample in the library as compared to the mean number of reads/sample in all samples across all six libraries. Then the number of reads for each ASV in each sample was corrected by this factor. (2) Dilution normalization: as a subset of the samples were diluted before sequencing based on the amplification levels as detected on agarose gel (see above) their ASV reads were multiplied by the dilution factor per sample to reflect their true original load. Table S2 presents the number of reads per ASV per sample after both data flooring and normalization.

Next, ASVs were aggregated based on UNITE classification, to species level when possible. ASVs that could not be classified to species level, were grouped together by the lowest known phylogenetic level and labeled "Other". Lastly, data were bubbled up

by summing up all reads in each taxonomic level to the taxonomies in the level above it (Table S3).

Decontamination

The negative control samples were then used to flag potential contaminant species (Figure 17C). Out of 456 species detected in the data (after flooring and normalization), 13 species unique to the negative control samples were removed from the dataset (Figure 17D). For an additional 63 species that were detected in both negative control samples and true samples, statistical testing was applied: (1) Fisher's exact test (on the floored normalized data converted to present/absent) was applied to check if a species was more prevalent in a specific condition versus the 191 extraction control samples. (2) Wilcoxon test (on the number of reads, without flooring, with library and dilution factor normalization) was applied to check if a species was more abundant in a specific condition versus the 191 extraction control samples. A species that had a p-value$\leq$0.05 and FDR$\leq$0.2 in at least one of these tests passed this filtering step for the condition. Next, the same tests were performed against the 104 paraffin control samples. Forty-two species (out of the 63 that were tested) passed both filtering steps in at least one condition. All of these 42 species, as well as the 380 species that did not appear in any of the 295 controls were considered part of the 'fungal world' that was used for all downstream analysis. The same filtering steps were also performed for each of the taxonomic levels (table S4). Note that only fungi and bacteria that passed the filtering steps in at least one of the tumor types were included in most of the analysis in this work.

Mock community

A mock community of 17 fungal species was generated to validate the ITS2 experimental procedure and assess the success of detecting different fungi (Table 5). DNA from all fungi was extracted using MasterPure Yeast DNA Purification Kit (Epicentre, MPY 80200). Equal amounts of DNA from each of the fungal species were mixed together and then 0.00032ng DNA of the mix was spiked into 100ng of human DNA (extracted from the HS-5 human fibroblast cell line (ATCC# CRL-11882)). ITS2 amplification and sequencing was done as described above.

Technical repeats of tumor and NAT samples

To assess the reproducibility of our technical and computational pipeline we repeated the ITS2 amplification and sequencing three times, for 88 human tumor or NAT samples. Samples were chosen to represent both high read samples and low read samples. ITS2 amplification and sequencing was done as described above. Only one sequencing result from each triplicate (the one with the highest amount of reads) was used for all other analyses that were subsequently done.

Construction and analysis of the multi-domain interaction networks

The following analysis was performed with MATLAB version 2019b with the Statistics and Machine Learning Toolbox. To construct the network, we first chose a taxonomic level for the fungi and bacteria. We then construct three different networks of interaction for each tumor type, fungus-to-fungus (FF) network, bacteria-to-bacteria (BB) network, and fungus-to-bacteria (FB), independently.

The relationship between each pair of taxa was calculated based on the presence/absence data, using the normalized mutual information (NMI) measure, which has been shown to perform as good or better than other ecological indicators of co-occurrence (37)

Given two vectors, X and Y, each with M discrete elements (corresponding to M samples), x_i and y_i (i=1…M) which can be equal to either 0 or 1, the NMI between them is defined as

$$NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}},$$

where $I(X,Y) = H(X) + H(Y) - H(X,Y)$ is the mutual information between X and Y, H(X) and H(Y) are the Shannon entropies of X and Y respectively, and H(X,Y) is the joint entropy of X and Y, i.e.,

$$H(X) = -P(x=0)\ log\ P(x=0) - P(x=1)\ log\ P(x=1)\ ,$$

$$H(Y) = -P(y=0)\ log\ P(y=0) - P(y=1)\ log\ P(y=1)\ ,$$

and

$$H(X,Y) = -\sum_{x_i \in (1,0)} \sum_{y_i \in (1,0)} P(x,y)\ log\ P(x,y)\ .$$

The NMI is bounded between 0 and 1, where 0 indicates no relationship between the presence/absence of taxon X and Y and 1 indicates maximal relationship (can mean that both always appear together or never appear together, i.e., it does not distinguish the sign of the relationship).

The p-value is calculated as the fraction of times a random reshuffling process of the taxon had outputted greater or equal NMI value to the original samples

$$p = \frac{\#NMI(X_{shuffle}, Y_{shuffle}) \geq NMI(X,Y)}{\#Shuffle\ realizations}.$$

To fairly compare the NMI values of random realizations, the shuffling is done in a weighted manner which preserves the total number of observed taxa in each sample. The weight w_i of each sample i is defined as

$$w_j = \frac{\#Observed\ taxa\ in\ sample\ i}{\#Total\ observed\ taxa\ in\ all\ the\ samples}$$

Then, the presence/absence of species X is randomly shuffled between the samples, with probability corresponding to the weight of each sample (i.e. the total number of fungi and the total number of bacteria present in each tumor was always kept as in the original data). The process is repeated 1000 times to calculate the p-value. We then perform BH FDR multiple comparison analysis on the p-values list of each tissue type and interaction type (FF, BB, and FB). Finally, a positive or negative sign of interaction was given to each pair of taxa according to a simple Pearson correlation. Only pairs with FDR≤0.25 were used in the figures (Figure 14).

5.8S real-time quantitative PCR (RT-qPCR)

RT-qPCR was performed on the 5.8S region of the fungal rDNA. The following primers were used: Forward primer (ITS3) - 5'-GCATCGATGAAGAACGCAGC-3' and reverse primer (ITS86R) - 5'- TTCAAAGATTCGATGATTCAC-3'. qPCR was performed on 40ng of DNA per sample (or the maximum available in 5ul). For extraction controls a volume of 5ul per sample was used. For empty paraffin controls a volume equal to the volume taken for the matching sample from the same block was

used. The PCR mix included 0.2uM of each primer, 5ul Kapa SYBR FAST qPCR Master Mix (2X) (Kapa Biosystems, #KK4605) and ultra pure water to a total volume of 10ul. PCR conditions used were 95°C 3min, (95°C 3sec, 58°C 20sec, 72°C 30sec) X40 cycles and included a dissociation curve at the end. ViiA 7 Real-Time PCR System (Applied Biosystems) was used for the qPCR. qPCR was performed in triplicates per sample and results were averaged across repeats. Fungal load was estimated by comparison to a standard curve created with Saccharomyces cerevisiae DNA that was spiked into human DNA.

Staining methods

Human tumor tissue microarrays (TMAs) were purchased from US Biomax and included over 400 cores representing the following tumor types: breast, lung, melanoma, ovary and PDAC. All TMA's were stained by H&E using standard protocol and serial sections from the same TMAs were used for the different stains (Figure 3A). All fungal antibodies were tested and their protocols calibrated on TMAs with known fungi in them that served as positive controls (Bio SB #BSB-0335-CS) (Figure 3B).

Modified Gomori Methenamine-Silver (GMS) Nitrate Stain

GMS (abcam #ab150671) was used for staining. Slides were deparaffinized and rehydrated as described above. Next they were washed in distilled water twice and incubated in chromic acid solution for 20 minutes. Slides were rinsed in tap water, and then washed in distilled water twice. Slides were then incubated in sodium bisulfite solution for 1 minute and then rinsed as before (1 tap water, 2 distilled water). Next slides were incubated in a pre-warmed GMS solution for 7 minutes at 60°C after which they were rinsed 4 times in distilled water and incubated in gold chloride solution for 30 seconds. Four additional distilled water rinses were performed followed by incubation in sodium thiosulfate solution for 2 minutes. Slides were next rinsed in tap water and 2 changes of distilled water. Next slides were stained with light-green solution for 2 minutes. Finally, slides were rinsed in absolute alcohol 3 times, left to dry and mounted with synthetic resin. For GMS protocol all tools used were plastic or glass (no metal-containing tools were used).

28S fungal fluorescence in-situ hybridization (FISH)

28S fungal FISH was performed with a mix of three fungal probes. 'D-205' probe: 5'-ATTCCCAAACAACTCGAC-3'; 'D-223' probe: 5'-CCACCCACTTAGAGCTGC-3'; and 'D-260' probe: 5'-TCGGTCTCTCGCCAATATT-3' (*57*), all conjugated to cy5 at the 5' end (IDT). Non-specific complement probes for each of the three probes as well as a mix of all probes together were tested on positive control tissues that were known to contain fungi in them, and found to have no background fluorescence (Figure 4A). For staining: slides were deparaffinized and rehydrated (Xylene for 10 minutes, Xylene for 5 minutes, 100% ethanol for 10 minutes X2, 96% ethanol for 10 minutes, 70% ethanol for 2-12 hours at 4°C). Slides were next rinsed in RNAse-free 2X SSC (Ambion #AM9765) for 10 minutes and proteinase K solution (10µg/ml in 2X SSC, Ambion #AM2546), pre-heated to 50°C was added to the slides. Slides were incubated for 10 minutes at 42°C. Slides were then rinsed twice with 2X SSC for 5 minutes each, followed by 2 rinses in wash buffer (2X SSC, 15% formamide (Ambion #AM9342)) for 5 minutes each. Next, slides were incubated overnight at 30ºC with a probe mix of 1 uM per probe in hybridization buffer (10% Dextran sulfate (Sigma #D8906), 15%formamide, 1mg/ml EcolitRNA (Sigma #R4251), 2X SSC, 0.02%BSA (Ambion #AM2616), 2mM vanadyl ribonucleoside (New England Biolabs #S1402S)). Slides were rinsed in wash buffer for 30 minutes at 30ºC followed by incubation in wash buffer with DAPI with a final concentration of 1ug/ml. Finally, slides were washed in 2XSSC, 10mM TRIS pH8 and 0.4% glucose and mounted with ProLong Gold Antifade Mountant (Life technologies #P36930).

Immunofluorescent staining

Slides were deparaffinized and rehydrated using the following protocol: Xylene for 10 minutes X 2, 100% ethanol for 5 minutes, 96% ethanol for 5 minutes, 70% ethanol for 5 minutes and 3 washes in PBS for 2 minutes each. Next endogenous peroxidase quenching was performed (1% H2O2, 0.185% HCl) for 30 min, followed by antigen retrieval using citric acid buffer (pH 6) for 10 minutes at 95ºC. slides were left to cool at room temperature and then washed 3 times in PBS. Blocking was done with 1% BSA and 0.2% Triton in PBS for 60 minutes at R/T. Slides were incubated with primary antibodies that were diluted using a staining buffer (2% horse serum, 0.2% Triton in PBS) overnight at 4ºC. The following antibodies were used: anti-1-3 b-glucan (abcam

#ab233743; 1:50), anti-CD45 (eBioscience #14-0459-82; 1:100), anti-CD68 (Invitrogen #MA5-12407; 1:50), anti-Aspergillus (Abcam ab20419; 1:100), anti-Dectin1 (Abcam ab140039; 1:200), anti-EFTU (Abcam ab90813; 1:300). Slides were washed in PBS for 2 minutes and secondary antibodies and DAPI (1ug/ml) diluted in staining buffer were added for 30 minutes at room temperature. The following secondary antibodies were used: Goat anti-Mouse IgG2b Cross-Adsorbed Secondary Antibody with Alexa Fluor 555 (Invitrogen #A21147; 1:200), Goat anti-Mouse IgG1 Cross-Adsorbed Secondary Antibody with Alexa Fluor 647 (Invitrogen #A21240; 1:200), Goat anti-Mouse IgG3 Cross-Adsorbed Secondary Antibody with Alexa Fluor 488 (Invitrogen #A21151, 1:200) and Donkey anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody with DyLight 755 (Invitrogen #SA5-10043, 1:100). Slides were washed twice in PBS and mounted with ProLong Gold Antifade Mountant (Life technologies #P36930).

Immunohistochemistry

Slides were stained by anti-fungal 1-3 β-glucan (abcam #ab233743; 1:100) and anti-bacterial LPS (Lipopolysaccharide Core, mAb WN1 222-5, 20 HycultBiotech #HM6011; 1:1000) or no primary antibody (negative control) with the automated slide stainer BOND RXm (Leica Biosystems) using the Bond polymer refine detection kit (Leica Biosystems #DS9800), according to the manufacturer's instructions. Acidic antigen retrieval was done by a 20 min heating step with the epitope retrieval solution 1 (Leica Biosystems #AR9961).

Imaging

Slides stained in all staining methods (IHC/IF/FISH/GMS/CFW) were scanned with the Pannoramic SCAN II automated slide scanner (3D HISTECH) at 40X.

Statistical analyses

Most of the downstream analysis and plots were performed with R version 4.03. Packages used in analysis include phyloseq 1.34.0, ggplot2 3.3.4, ggbeeswarm 0.6.0, ggrepel 0.9.1. VennDiagram 1.6.20, pheatmap 1.0.12, ggforce 0.3.3, ggpubr 0.4.0, RColorBrewer 1.1-2, proxy 0.4-26, reshape2 1.4.4, stringr 1.4.0, dplyr 1.0.7, purrr 0.3.4, readr 1.4.0, tidyr 1.1.3, tidyverse 1.3.1. Note that P values less than 2.2×10−16

are not reported by ggpubr, so P values less than this are listed as <2.2×10−16; it is not a range of P values.

Human RNA depletion protocol

Sample preparation

To test the depletion protocol mixes of human, fungal and bacterial RNA were used. Human HS5 cells were cultured in DMEM medium supplemented with xx ng/ml Amphotericin B (Sigma, A2942-20ML) to ensure no fungal contaminations and xx ng/ml pen/strep to ensure no bacterial contamination. RNA was extracted with the Perfect pure RNA cultured cell kit (5 prime, FP2302340) according to manufacturer's protocol. Fungal RNA was extracted from Saccharomyces cerevisiae grown over night in YPD suspension and bacterial RNA was extracted from Escherichia coli grown overnight in LB suspension to an OD of 0.5-1 with GeneJET RNA purification kit (Thermo Fisher, K0731) according to manufacturer's protocol. Human, fungal and bacterial RNA were mixed at a ratio of 100:1:1 cells respectively (based on RNA content per cell from bionumbers: http://bionumbers.hms.harvard.edu/).

Probe preparation

Probes were prepared from human HS5 cells. HS5 cells were grown and RNA was extracted as described above. RNA was fragmented using 4ul FastAP buffer (10x) (Thermo Scientific, EF0651) and 1µg of RNA in a total of 20µl. Mix was incubated for 4 minutes at 94°C in a pre-heated thermal cycler. Next samples were placed on ice. Sample was cleaned and size selection was performed by RNAClean XP beads (Beckman Coulter, A63987). Beads were resuspended by vortexing and X2 volume of beads was added and mixed with sample by pipetting. Beads and RNA were incubated at room temperature for 15 minutes and placed on a magnet for 5 minutes until solution was clear, solution was discarded and 200µl of fresh 75% ethanol was added for 30 seconds X2 repeats. Solution was discarded and beads were left to dry for ~5 minutes. Tube was removed from the magnet and RNA was eluted by adding 32µl of nuclease free water. Mixed well by pipetting 10 times and incubated for 2 minutes at room temperature. Tube was then placed on the magnet and after 5 minutes, solution was transferred to a clean tube.

cDNA was prepared from the RNA mix using MultiScribe Reverse Transcriptase (Thermo Fisher, 4311235) according to manufacturer's protocol.

After cDNA synthesis RNA was degraded by adding 10% reaction volume of 1N NaOH and incubation at 70°C for 12 minutes. Reaction was neutralized with 20% of initial volume of 0.5M acetic acid.

Reaction was cleaned with RNAClean XP beads as described above but with 70% ethanol and final elution in 22µl. Samples were tested for quality by Tapestation after each stage of the process.

Depletion protocol

For depletion testing 500ng of RNA mix (human:yeast:bacteria) was used. RNA was first degraded using FastAP (10X) and cleaned up in the same way as performed for probe fragmentation but for 3 minutes. Next probes were hybridized to the RNA mix. Since there was a problem with the quantification of the probe concentration, two concentration were tested. Max concentration in which the maximum volume of probe was used (12µl) and Min concentration in which half of it was used (6µl), an additional NC (negative control) with no probe was performed. Depletion reaction was performed as follows: 1ul of RNA (10ng/µl) was mixed with 2ul probe hybridization buffer (100mM Tris-HCL, 200mM NaCl, pH 7.4), ultra pure water and probes at a final volume of 15µl. Depletion mix was incubated in a PCR machine with 2 ramp steps: for 95°C for 2 minutes, 95-80°C for 2 minutes and 3 seconds lowering the temperature by 0.1°C/second, 80-37°C for 35 minutes and 50 seconds lowering the temperature by 0.02°C/second, 37°C for 5 minutes. Samples were spun down and placed on ice. Next RNA in RNA-cDNA strands was digested with RNase H. RNase mix was prepared on ice: 2µl of NEBNext RNase H (NEB, M0297S), 2µl of RNase H Reaction Buffer, 1µl of nuclease free water per reaction. RNase mix was added to the sample and mixed by pipetting. Mix was spun down and incubated at 37°C for 60 minutes. Mix was spun down and placed on ice. Next cDNA was digested by DNase I. DNase mix was prepared on ice: 2.5µl of DNase I (RNase-free) (NEB, M0303S), 5µl of DNase I Reaction Buffer, and 22.5 nuclease free water per reaction. DNase mix was added to the sample and mixed by pipetting. Mix was spun down and incubated at 37°C for 60 minutes. Mix was spun down and placed on ice. Next samples were cleaned with RNA Clean XP

beads as described above but with 85% ethanol and eluted in 18ul. Samples were tested for quality by tapestation after several steps.

Depleted RNA samples were next used for cDNA synthesis using MultiScribe Reverse Transcriptase (Thermo Fisher, 4311235) according to manufacturer's protocol.

qPCR testing of depletion successes

To test depletion protocol success qPCR was performed. Primers used are in the table 6 below:

Gene	Organism	Primer name	Sequence (5' -> 3')
RNA18S	Human	rt-c-18S-F	CATTCGTATTGCGCCGCTA
		rt-c-18S-R	CGACGGTATCTGATCGTCT
18S	S. Cerevisiae	ySC-18S-F	TGGCGAACCAGGACTTTTAC
		ySC-18S-R	CCGACCGTCCCTATTAATCAT
16S	E. Coli	bEC-16S-F	ACCCACTCCCATGGTGTGA
		bEC-16S-R	GAATGCCACGGTGAATACGTT
PABPC1	Human	PABPC1-F	GCACAAGTTTCTTTTCATGGTCC
		PABPC1-R	AGTCACTCCGTTCTAAGGTTGA

qPCR was performed on RNA and cDNA samples after depletion. RNA samples were used to test if cDNA probe degradation was complete. qPCR was performed with KAPA SYBR FAST qPCR Master Mix (2X) Kit (KAPA Biosystems, KM4103) according to manufacturors protocol, in a total volume of 10µl. qPCR conditions used were 95°C for 3 minutes, (95°C for 3 seconds, 60°C for 30 seconds)X40, 72°C for 30 seconds.

TCGA, Hopkins and UCSD cohort methods

TCGA cohort: Data accession

All TCGA sequence data were accessed via the Cancer Genomics Cloud (CGC) as sponsored by SevenBridges (https://cgc.sbgenomics.com) (*86*) after obtaining data access from the TCGA Data Access Committee through dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). Details of how TCGA samples were acquired and processed are comprehensively described elsewhere (*87*), and SOPs for TCGA sample processing are available in the NCI Biospecimen Research Database (https://brd.nci.nih.gov/brd/sop-compendium/show/701). Metadata for these

patients were previously published and originally compiled using SevenBridges's metadata ontology (*18*).

Hopkins and UCSD cohorts: Data accessions

Raw BAM files for the Hopkins plasma cohort were accessed through the European Genome-Phenome Archive (EGA) under Study ID EGAS00001003611 with prior data access approval. These files were previously analyzed for host-centric, fragmentomic cancer diagnostics by Cristiano *et al*. (*60*). Raw BAM files for the UCSD cohort were internally available after Poore and Kopylova *et al*. (*18*) previously published them using bacterial-centric analyses, and host-depleted versions of the files are publicly available on European Nucleotide Archive (ENA) with the following accession IDs: ERP119598 (UCSD HIV-negative controls), ERP119596 (UCSD prostate cancer), and ERP119597 (UCSD lung cancer and melanoma).

TCGA, Hopkins, and UCSD cohorts: Library preparation and sequencing

Library preparation and sequencing methods of TCGA were described in detail by Hoadley *et al*. (*87*), and primarily employed QIAGEN products for multi-analyte (DNA, RNA) extraction and Illumina platform sequencing. Sample processing and sequencing of the Hopkins cohort was described in detail by Cristiano *et al*. (*60*), and, briefly, performed cell-free plasma DNA extraction using the Qiagen Circulating Nucleic Acids Kit, non-fragmented library preparation using a modified protocol of the NEBNext DNA Library Prep Kit for Illumina, and sequencing with 100-bp paired-end runs on the Illumina HiSeq machines (*60*). Sample processing and sequencing of the UCSD cohort was described in detail by Poore and Kopylova *et al*. (*18*), and, briefly, performed cell-free DNA extraction using the Qiagen Circulating Nucleic Acids Kit, library preparation using the KAPA HyperPlus Kit (Kapa Biosystems), and paired-end 2×150-bp sequencing on an Illumina NovaSeq 6000 instrument (S4 flow cell).

TCGA, Hopkins, and UCSD cohorts: Bioinformatic processing

Determining read counts in TCGA

Total and mapped read counts were calculated using SAMtools's idxstats function (v. 1.11) (*88*), which was wrapped in a Docker container and applied to all available TCGA BAM files on the CGC as an "app." The app was then run in parallel across files using

Amazon Web Services (AWS) as a backend using 8 cores per file. Total read counts were extracted from the resultant idxstats output files using *awk '{s+=$3+$4} END {print s}'* and mapped read counts were extracted using *awk '{s+=$3} END {print s}'*. Unmapped read counts were determined via the subtraction of mapped from total. Microbial read counts were derived by summing all genome-level microbial hits against the "rep200" database (see below for more details). Similarly, fungal-specific or bacterial-specific counts were determined by summing all genome-level microbial hits against the rep200 database within those domains.

Host depletion of WGS and RNA-Seq data

Previous efforts to mine host genomic or transcriptomic information for microbial nucleic acids relied on extracting unaligned, "non-human" reads from pre-aligned BAM files, followed by mapping those reads against a database of microbial genomes (*18*) Since TCGA samples were collected during a decade (2006-2016), the human genome references used for BAM file generation changed over time, and uniform realignments were not performed until very recently (*89*). Although this was not detected to be a problem by Poore and Kopylova *et al*. (*18*) for bacterial-centric analyses, we wanted to uniformly host deplete and further quality control all TCGA files prior to multi-domain mapping and metagenome assembly. Thus we designed, optimized (for speed), and Dockerized a uniform host depletion pipeline using a combination of SAMtools (v. 1.11) (*88*), Minimap2 (v. 2.17-r941) (*90*), and fastp (*91*) capable of being run on any high performance compute system.

Read pairs were subsequently discarded if either mate mapped to the GRCh38.p7 human genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.33/) or the Phi X 174 viral genome. Reads were also discarded if less than 45 base pairs in length or if they exhibited poor base quality (using fastp default parameters). Specifically, the following command was run, where $cpus and $db denote the number of compute cores and a precomputed Minimap2 reference database (as a .mmi file), respectively:

samtools view -f 4 -O BAM $in_dir/$filename |

samtools bam2fq - |

fastp -l 45 --stdin -w $cpus --stdout --interleaved_in |

minimap2 -ax sr -t $cpus $db - |

samtools fastq -@ $cpus -f 12 -F 256 - -1 $out_dir/$base_name.R1.trimmed.fastq.gz -2 $out_dir/$base_name.R2.trimmed.fastq.gz

The final line outputs forward ("R1") and reverse ("R2") fastq files. Sometimes, due to cloud computing constraints, the first line of the command (samtools -f 4) was done separately from the remaining lines, which were consistently run together. Typical compute time per file for the host depletion and read extraction ranged from several minutes to a few hours using 8-16 cores and ~100 GB of RAM.

We note that this additional host depletion reduced the number of total files available for the TCGA mycobiome analysis when files resulted in 0 non-human reads. Specifically, 77 WGS files and 2530 RNA-Seq files had 0 non-human reads after additional host depletion and could not be used for shotgun metagenomic or metatranscriptomic microbial assignments. Another 16 files repeatedly failed the host depletion pipeline and could not be used. Overall, this reduced the number of files available for the TCGA mycobiome analysis compared to our previous bacteriome-centric analysis (*18*).

Shotgun metagenomic and metatranscriptomic microbial assignments

Host depleted and quality controlled output fastq files were then uploaded to Qiita web server (*92*) for per-sample metagenomic or metatranscriptomic microbial classification. Qiita offers a graphical user interface that facilitates shotgun metatranscriptomic and/or metagenomic analyses using direct genome alignments based on Woltka v0.1.1 (https://github.com/qiyunzhu/woltka) (*93*) against Qiita's concomitant "rep200" multi-domain database. "Rep200" corresponds to RefSeq release 200 (built as of May 14, 2020), which comprises 11,955 genomes with the following taxa: 419 archaea; 11,080 bacteria; 320 fungi; 88 protozoa; 48 viruses (https://qiita.ucsd.edu/static/doc/html/processingdata/processing-recommendations.html#reference-databases). We note that the only other database used for Qiita metagenomics or metatranscriptomics (Web of Life, WoL) does not include fungi. Direct genome alignments against rep200 were run using Bowtie2 v2.4.1 (*94*) as the backend. This process is equivalent to a Bowtie2 run with the following parameters:

--very-sensitive -k 16 --np 1 --mp "1,1" --rdg "0,1" --rfg "0,1" --score-min "L,0,-0.05"

The sequence alignment is treated as a mapping from queries (sequencing data) to subjects (microbial reference genomes). Reads mapped to a microbial reference genome are counted as hits, such that the resultant feature table comprises samples (rows) by microbial genome IDs (columns) and concomitant abundances. These microbial genome IDs (named "operational genomic units" or OGUs) provide a shotgun metagenomic equivalent to ASVs in 16*S* rRNA gene amplicon sequencing data (*93*). Of note, in the case that one sequence is mapped to multiple genomes by Bowtie2 (up to 16), each genome is counted 1 / k times, where k is the number of genomes to which this sequence is mapped. The frequencies of individual genomes were then summed after the entire alignment was processed, and rounded to the nearest even integer, thereby making the sum of OGU frequencies per sample is nearly equal (considering rounding) to the number of aligned sequences in the dataset. The resultant count matrix was saved as a biom file for downstream analyses. This process was repeated for the TCGA, Hopkins, and UCSD cohorts, with separate Qiita projects under the following study IDs: 4736 (TCGA WGS), 13767 (TCGA RNA-Seq), 13984 (Hopkins), 12667 (UCSD HIV-free controls); 12691 (UCSD prostate cancer); 12692 (UCSD lung cancer and melanoma).

TCGA cohort: α and β diversity calculations

Alpha diversity calculations

Raw fungal count data from primary tumors was subset to each TCGA WGS sequencing center and processed using QIIME 2 (version qiime2-2020.2) (*95*) calculate richness and shannon alpha diversity per center. Rarefaction amounts were determined by the fungal read count distribution per TCGA sequencing center, and a common value of 5000 reads/sample was identified among 4 of the 5 WGS sequencing centers as being approximately the first quartile of reads/sample—Broad Institute WGS samples excepted, and 2000 reads/sample approximately represented the first quartile and was used.

Alpha diversity fungi-bacteria correlations

Multi-domain TCGA alpha diversity was calculated using the following procedure: (1) Subset to WGS primary tumor samples; (2) rarefy the entire WGS rep200 table to 130,000 reads/sample (approximately the first quartile of the read/sample distribution) using phyloseq (v. 1.38.0) (*96*); (3) separate fungal and bacterial features into two separate tables; (4) calculate richness among both fungal and bacterial rarefied tables; (5) correlate, using Spearman correlations, the paired fungal and bacterial richnesses (Figure 7C). We also attempted this procedure with the modification that fungal and bacterial feature tables were independently rarefied, but we found that this version caused microbial richness to weakly but still significantly, positively correlate with the sample library sizes (data not shown), so it was discarded in favor of the 'global' table rarefaction.

β-diversity calculations

Given the limited fungal reads/sample, we desired to perform β-diversity without rarefying using a method we previously published named robust Aitchison PCA (RPCA, also called DEICODE) (*97*) (https://library.qiime2.org/plugins/deicode/19/). DEICODE has a QIIME 2 plugin that was used on the raw fungal count data in primary tumors subset to each TCGA WGS sequencing center with the following parameters: {--p-min-feature-count 10, --p-min-sample-count 500}. The resultant biplots were visualized using EMPeror (*98*) and the QIIME 2 plugin for ADONIS (i.e., PERMANOVA) was used to estimate the significance and explained R^2 of cancer type with the DEICODE distance matrix.

To compare tumor vs. NAT samples in TCGA, we performed the following analyses:

Analysis #1: (1) Rescale Voom-SNM batch corrected pan-cancer data into counts using a scalar of 10^4; (2) calculate relative abundances using the batch corrected counts; (3) average fungal relative abundances across disease type-sample type groups (e.g., "Breast Invasive Carcinoma NAT"); (4) calculate Bray-Curtis dissimilarity on the averaged relative abundances; (5) plot using a principal coordinates analysis using cancer types also found in the Weizmann cohort and with at least 10 tumors and NATs available in TCGA. Sample counts: breast NAT, n=100; breast tumor, n=978; colorectal NAT, n=72; colorectal tumor, n=526; Lung NAT, n=194; Lung tumor,

n=1068; Ovarian NAT, n=10; Ovarian tumor, n=683. Note that “lung” combines TCGA projects LUAD and LUSC and that “colorectal” combines TCGA projects COAD and READ." (Figure 8E).

Analysis #2: We performed a PERMANOVA analysis within each disease type for sample-type for both Aitchison and Bray-Curtis distances on the full sample set of relative abundances and found that no disease type significantly differs between tumor and NAT after accounting for multiple testing correction (table S9).

TCGA, Hopkins, and UCSD cohorts: Decontamination

TCGA decontamination

Although TCGA protocols did not include contamination controls during the processing of their samples, we showed that *in silico* methods could be used to decontaminate the TCGA bacteriome (*18*). The fundamental principle of these methods is that consistent negative correlations exist for external (e.g., reagent, environmental) contaminating taxa between their read fractions and analyte (DNA or RNA) concentrations (*99*). A published tool named *decontam* (https://github.com/benjjneb/decontam) (version 1.14.0) (*99*) wraps the method into an R package and function based on two underlying mathematical assumptions: (i) the contaminants are added in uniform amounts across samples; and (ii) the amount of contaminant DNA or RNA is small relative to the true sample DNA or RNA (microbial or host). Since per-sample DNA and RNA concentrations are available in TCGA metadata, they can be used to indicate putative contaminating taxa. Importantly, though, our past analyses (*18*) demonstrated that too stringent of an *in silico* decontamination threshold actually removes flora known to be associated with a given body site (e.g., too stringent decontamination of NAT colon tissues in TCGA dissociates it from normally-associated fecal material). Additionally, there are difficulties of strict filtering with taxa that are known commensals and/or pathogens but also can be contaminants in certain contexts, even at the species level (e.g., *Malassezia restricta*, a skin fungus). Thus, in our mycobiome analyses, we sought a balance between strict filtering, allowance of known commensals/pathogens, inclusion of WIS-identified (this study) or HMP-identified (*58*) fungi, and inclusion of fungi of unknown significance that may be related to cancer biology.

Decontamination was thus broken into two steps: (i) Statistical decontamination via *decontam* using per-sample DNA or RNA concentrations and read fractions across plate-center batches (see below); and (ii) manual curation, comparison against WIS-identified and HMP-identified fungi, and literature review prior to making final determinations.

Step #1: TCGA sample identifiers (e.g., "TCGA-02-0001-01C-01D-0182-01") denoted the sequencing center and plate within that center upon which the sample was run (for details, see https://docs.gdc.cancer.gov/Encyclopedia/pages/ TCGA_Barcode/). These barcodes were used to extract all sequencing plate-center combinations using the last two sets of integers (e.g., "0182-01" is plate 182 from center 1). We previously found the plate-center method to work well on TCGA bacterial data, as it removed many likely contaminants while retaining several known commensals and pathogens (*18*). Since *decontam* effectively performs a regression analysis to determine if a taxon is a contaminant, we required ≥10 samples per plate-center batch, retaining 325 total plate-center batches among samples positive for fungi. *Decontam* was then run in "frequency" mode, identifying putative contaminants using TCGA sample aliquot concentrations, a default P* stringency threshold of 0.1, and the default batch.combine="minimum" parameter, such that a taxon was removed if identified in any one of the 325 plate-center batches as a contaminant. This analysis identified 71 putative contaminants out of 319 total fungi with ≥1 reads identified during direct genome alignments. Table S6 summarizes the *decontam* output and contaminant predictions.

Step #2: All 319 fungal taxa found in TCGA were cross-referenced against species identified in the WIS tumor mycobiome cohort (this study), the HMP gut mycobiome cohort (*58*), and 131 other papers in the literature (table S6). This comprehensive literature survey informed the final decontamination decisions. Specifically, the following decision making process was applied: (i) Any fungal specie identified in the WIS tumor mycobiome cohort or HMP gut mycobiome cohort was retained; (ii) any fungal specie known in the literature to have caused a clinically pathogenic infection or be a human commensal was retained; (iii) any fungal specie with evidence of no known human association was discarded; (iv) any specie that had little evidence for or against human associations (i.e., "unknown" human associations) had their fate decided by the plate-center *decontam* predictions. This process ultimately discarded 95 species (29.8%

of total) as contaminants, comprising 2.2% of total reads, and retained 224 species as non-contaminants (table S6).

Hopkins cohort decontamination

The Hopkins plasma cohort was originally collected to examine host-centric fragmentomic diagnostics (*60*) and did not employ contamination control samples. Since the TCGA contamination analysis thoroughly covered 319 out of 320 total fungi in the rep200 database, the contamination decisions from TCGA based on the WIS cohort, HMP gut mycobiome cohort (*58*), and 131 other papers were applied to the Hopkins cohort. The Hopkins cohort began with 296 identified fungal species and after decontamination retained 209 fungal species (29.4% removed).

UCSD cohort decontamination

The UCSD plasma cohort was designed to include positive and negative contamination control samples (*18*). Positive controls included 26 samples of serially diluted *Aliivibrio fischeri* (bacteria), which were previously analyzed (*18*), while negative controls included 15 blank DNA extraction samples and 11 blank library preparation samples. All control and biological samples were run on a single sequencing plate at a single time, as described previously (*18*). Decontamination was performed using *decontam* in (i) "prevalence" mode with P*=0.5 among blanks and biological samples, and in (ii) "frequency" mode using the default P*=0.1 (also used in TCGA) with DNA concentrations. Importantly, for "prevalence" mode, P*=0.5 will flag taxa as contaminants if they are more prevalent in negative controls than in biological samples. These were run separately because several of the blanks had zero or otherwise undetectable DNA concentrations, which are compatible with "prevalence" filtering but not "frequency" filtering. "Prevalence" filtering flagged 30 out of 227 (13.2%) identified fungi while "frequency" filtering identified 4 out of 227 identified fungi (1.8%), or 32 unique total fungi (14.1%). These putative contaminants were then compared against the comprehensive TCGA decontamination analysis and guided the decision of any "unknown" human associated fungi. As with TCGA and the Hopkins cohorts, fungi matching the WIS cohort, HMP gut mycobiome cohort (*58*), or with known pathogenic/commensal associations were retained whereas those with evidence

against human associations were removed. This ultimately left 215 decontaminated fungi for analysis in the UCSD cohort.

TCGA cohort: Co-occurrence analyses with MMvec

In order to explore the fungal genera identified in the controlled amplicon based sequencing at a large scale, the TCGA metagenomic dataset count table was group summed to the genus level and matched to genera in the WIS amplicon data. This process was then repeated for the bacterial data, so that both tables were operating at the same taxonomic level and only contained WIS-overlapping features (table S10). TCGA immune compositions were obtained from Thorsson *et al*. (*40*), who derived them using CIBERSORT (*69*) on TCGA RNA-Seq samples. Note that TCGA performed combined RNA-Seq and WGS on many samples, enabling usage of the WGS data to inform microbial composition and paired RNA-Seq data to inform immune cell composition. RNA-Seq data was not used for co-occurrence analyses due to (i) much lower read microbial depths and (ii) bias in the bacterial data due to polyA selection as noted in TCGA SOPs. In this case, TCGA patient identifiers published by Thorsson *et al*. (*68*) were used to match immune cell compositions to microbial data.

MMvec (v. 1.0.6) (*70*) was optimized between each data modality (i.e., bacteria, fungi, and immune cell composition) within each submission center (Harvard Medical School, Baylor College of Medicine, and MD Anderson) to (i) avoid center effects and (ii) produce a minimized cross-validation (CV) error, log-loss, and a maximized Q-squared (1 − model coefficient of variation [CV] / null model CV) values. Note that a Q-squared value > 0 ensures a good model fit. Training and test labels were produced across all tables stratified by cancer type. Each model had the following optimized parameters: 2e3 to 5e3 iterations, batch size of one fourth the training tables numbed of features, number of epochs as (# of iterations * batch size) / total reads in the training table, latent dimension of 3, and all other parameters were set to default. The null model operated on the exact same training/test set and parameters with the exception of the latent dimension set to zero. All models produced between all data modalities and submission centers had Q-squared values greater than zero, verifying their fits.

To explore co-occurrence clusters between all data modalities, MMvec conditional probabilities were z-score transformed along the first axis (i.e., across columns of the

MMvec output, as done elsewhere (*100*)). In order to minimize the effect of the TCGA submission center, we explored only those features with consistent co-occurrences across TCGA submission centers—defined as features whose median co-occurrence values were less than the standard mean error (SEM) of their co-occurrence values across centers. Next, the median of these filtered features were taken across all submission centers. To explore the co-occurrence clustering and define subtypes across modalities, hierarchical clustering was performed through Scipy's (v. 1.3.0) hierarchy linkage function (*101*) via Seaborn's (v. 0.11.2) (*102*) clustermap plotting function. Three fungi-driven "mycotypes," or subtypes, were identified across the highest partition of linkages on the immune co-occurrences. These subtypes were defined as follows: F1 (*Malassezia*, *Ramularia*, and *Trichosporon*), F2 (*Candida*, *Aspergillus*), and F3 (*Tilletiopsis*, *Penicillium*, *Cryptococcus*, *Puccinia*, *Agaricus*, *Alternaria*, *Phialocephala*, *Fusarium*, *Hyphopichia*, *Exophiala*, *Stereum*, *Colletotrichum*, *Dissoconium*, *Aureobasidium*, *Talaromyces*, *Cutaneotrichosporon*, *Yarrowia*, and *Trichoderma*). The immune cells and bacterial genera associated with each mycotype were then defined by their within-linkage-cluster maximum co-occurrences.

TCGA and UCSD cohorts: Batch correction

TCGA data was collected across a decade at multiple sequencing centers, sequencing platforms, and experimental strategies (WGS vs. RNA-Seq) among other technical variables. Fortunately, strict SOPs limited other forms of variation between centers. Our previous analyses on the TCGA bacteriome suggested that the largest technical factors were (in order from most to least) experimental strategy, sequencing center, and sequencing platform. Collectively, these factors explained 95.9% of the variability in bacterial data (*18*, *103*) using principal variance components analysis (PVCA) and necessitated batch correction prior to pan-cancer analyses. We found a similar effect within the fungal data, which motivated subsetting all samples to Illumina HiSeq platform, comprising 97% of samples (see Supplementary Note), and performing batch correction on the experimental strategy and sequencing center, which explained 49% and 30% of variance, respectively, using PVCA (data not shown). Batch correction was applied using the combination of Voom and SNM, as done previously (*18*, *103*, *104*). Briefly, Voom converts discrete counts to pseudo-normally distributed ("microarray-like") data (*103*), which is then used by SNM to iteratively remove batch effects in a supervised manner (*104*), such that biological signal is not removed while technical

variance is removed. PVCA was used before and after batch correction (*105*), as recommended by the National Institute of Environmental Health Sciences (NIEHS) (https://www.niehs.nih.gov/research/resources/software/biostatistics/pvca/index.cfm). We set the single tunable parameter for PVCA (the percentage of variance explained to obtain a number of PCA components) to 80%, based on NIEHS's recommendation of 60–90% and our past analyses (*18*).

For Voom and SNM, the biological variable was sample type (e.g., tumor, NAT, blood) for TCGA and disease type for the UCSD cohort, both as done previously (*18*). During exploratory analyses of immunotherapy response in the UCSD cohort (data not shown), the patient treatment response status was included as another biological variable in addition to disease type. We briefly but importantly note that SNM was designed for all possible biological variables to be included, including those that would later be examined using differential expression/abundance testing (*104*). For technical factors, the TCGA cohort used experimental strategy and sequencing center, whereas age and sex were used for the UCSD cohort, both also done previously on the same cohorts (*18*). As with bacterial-centric data, PVCA on TCGA before and after batch correction on mycobiome data showed remarkable reduction in technical variable variance up to 20.4-fold while retaining or increasing (i.e., improving the signal-to-noise ratio) biological variable variance up to 7.7-fold (data not shown).

When subsetting feature sets to those with (i) 34 WIS-overlapping fungal species, (ii) 31 fungal species with ≥1% aggregate genome coverage, (iii) the top 20 Hopkins-associated fungi (table S11) , or (iv) overlapping WIS fungi and bacteria (approximately 300 species depending on the intersected dataset), the raw count data were first subset followed by Voom-SNM. This means that batch correction occurred independently on each smaller feature set prior to downstream machine learning. Performing PVCA on each of these feature sets before and after batch correction frequently showed similar reductions in technical variable variance and maintenance or increases in biological variable variance (data not shown).

All cohorts: Machine learning methods

Note of caution when interpreting AUROC and AUPR values

It is common to estimate ML performance using area under ROC (AUROC) and PR (AUPR) curves; however, there are important differences between them, as they measure different aspects of discrimination and have different null values. Specifically AUROC on a model that performs as good as random coin flipping would be approximately 50%, and this calculation takes into account both true positives and true negatives. However, the AUPR of a model that has null performance would actually have differing null areas depending on the underlying prevalence of the positive class, and the calculation does not take into account true negatives. For example, TCGA contains many more tumor samples than NAT samples, and we model tumor samples as the positive class since it represents an active diagnosis of cancerous tissue. A model that performs randomly on tumor vs. NAT discrimination would have an AUROC of ~50% but a much higher AUPR (e.g., in a hypothetical case, if we had 90 tumors and 10 NAT samples, the null AUPR would be 90%). Furthermore, the calculations of precision and recall on the resultant predictions would not take into account how many samples were true negatives (i.e., those predicted to be NAT and indeed being NAT). Both of these can make interpretation of AUPR difficult, especially when compared to one-cancer-type-versus-all-others ML models, where the prevalence of the positive class (cancer type of interest), and thus null AUPR, is often in the range of 1-10%. Nonetheless, it is common to advocate for measuring AUPR in addition to AUROC when classes are imbalanced, since large class imbalances in certain circumstances can artificially raise AUROCs. Thus, for these analyses, we have consistently calculated both and indicated the null AUROCs and null AUPRs on most ML performance plots, and we continue to caution that for analyses where true negatives are important AUROCs may be more appropriate to examine.

ML of individual cancer types versus each other or controls

We previously published ML on the TCGA bacteriome using stratified 70% training, 30% holdout testing splits (*18*) across all cancer types. While suitable for the large number of ML models being built and tested within and between cancer types in TCGA, this strategy did not provide information of performance error ranges. We thus decided

to modify the strategy for the mycobiome analyses in such a way to provide both the performance estimate and a confidence interval for that performance across each cancer type without largely increasing compute times for each model. Specifically, for each model, we performed 10-fold cross validation using gradient boosting models (GBMs) with ten stratified 10% holdouts (i.e., the prediction class proportions are similar in train/test, such that if the entire dataset was 10 positive class and 90 negative class, then each k^{th} holdout would have 1 positive class sample and 9 negative class samples). ROC and PR curves and areas were calculated for each 10% holdout test set, such that ten sets of two-class discriminatory performance—effectively ten sets of 90% training-10% testing—were obtained for each model. These performance estimates were then aggregated for each model to calculate the 95% confidence intervals of performance. One other key difference between this and our previous approach (*18*) is that the hyperparameter grid search was removed in favor of a fixed GBM grid with the following parameters: {n.trees=150, interaction.depth=3, shrinkage=0.1, n.minobsinnode=1}. We note that these parameters were possible in our past TCGA analysis (*18*) and were equal to those used in the host-centric analyses of the Hopkins cohort (*60*) (https://github.com/cancer-genomics/delfi_scripts/blob/master/06-gbm_full.r). Equal to our last approach (*18*), we also up-sampled the minority class in cases of class imbalance while requiring ≥20 samples in the minority class to help the model generalize. We also centered and scaled the data prior to ML model building when using Voom-SNM batch corrected data; however, when using raw count data, we only removed zero variance features prior to the ML model building. This approach of the ML was then rapidly iterated on TCGA, WIS, Hopkins, and UCSD cohorts, collectively representing hundreds of models and thousands of train-test splits. We also note that in the case of WIS data, all filtered fungal or bacterial hits were used regardless of taxonomic rank (i.e., "free rank" data), based on empirical performance benefits, whereas ML in TCGA, UCSD, or Hopkins was performed with data summarized to a single taxonomic level (e.g., species, genus).

Multi-class ML in TCGA using raw data

During validation analyses on raw TCGA count data (see Supplementary Note), we noticed that independently training ML models on two stratified TCGA halves and subsequently testing on the other half provided highly concordant performance (data

not shown). (Note that stratified samples were based on sequencing center, sample type, and disease type and that experimental strategy was covered since 7 of 8 sequencing centers only performed WGS or RNA-Seq, and the one that did both [Broad] processed 83% of samples with WGS only.) This motivated testing whether multi-class machine learning was possible using stratified train-test splits, again by sequencing center and disease type. Since experimental strategy had the largest batch effect (see "TCGA cohort: Batch correction" section above), we conservatively used WGS-only samples to ensure that multi-class ML performance would not be affected by WGS vs. RNA-Seq variability while continuing to stratify splits by sequencing center, sample type, and disease type. This type of ML came up in two circumstances: (1) Comparing the pan-cancer ML performance in TCGA of WIS-overlapping fungal species vs. equal sized feature sets of non-WIS-overlapping features in tumor tissue and blood, and (2) comparing the relative pan-cancer performance in TCGA of WIS-overlapping fungi, bacteria, or both. Details of these are provided below, and as above, we note that up-sampling the minority class and removing zero variance samples were continued here.

Case #1: A total of 34 fungal species overlapped between TCGA and WIS cohorts. To test whether these features were more informative when discriminating between cancer types versus similarly sized feature sets, we did the following: (1) Randomly sample 34 non-WIS-overlapping fungal species; (2) create stratified 70% train-30% holdout test sets among WGS samples; (3) train two pan-cancer ML models using multi-class classification on the 70% stratified training set, one using WIS fungi and another using non-WIS fungi; (4) test both trained models on the holdout 30% stratified test set; (5) calculate average performance (AUROC, AUPR) across all one-cancer-type-versus-all-others comparisons after applying each model to the test set; (6) repeat steps 1-6 for a total of 100 times; (7) repeat for both primary tumor and blood derived normal samples. The resultant performance indeed suggested that WIS-overlapping fungi provided better pan-cancer discriminatory performance (Figure 9B).

Case #2: To test whether adding fungal to bacterial information would improve pan-cancer discrimination, we did the same procedure as Case #1 with the following differences: (1) Three feature sets were used, consisting of WIS-overlapping fungi, WIS-overlapping bacteria, and both WIS-overlapping fungi and bacteria; (2) all three feature sets were used to train and test ML models based on the stratified 70% training and 30% holdout test sets. We note that WIS-overlapping features were used for these

analyses because they represented the most confident species calls identified in two international cohorts. The resultant performance indeed suggested that combining fungal and bacterial information synergistically provided better pan-cancer discriminatory performance (Figure 9C).

Hopkins and UCSD pan-cancer analyses

Cristiano *et al*. (*60*) originally benchmarked the performance of host-centric, fragmentomic, pan-cancer diagnostics using GBM ML models based on 10-fold cross-validation repeated 10 times using the following model hyperparameters: {n.trees=150, interaction.depth=3, shrinkage=0.1, n.minobsinnode=1}. Notably, the only major ML difference between their method and ours (described above) was that we did not repeat the 10-fold cross validation ten-times. Thus, to directly compare our pan-cancer performance on the Hopkins cohort with their previously published results, we implemented an approach to repeat the 10-fold cross-validation ten-times, such that the ten iterations of performance measurement were done on the aggregated predictions. In other words, the first iteration of this method created 10 sets of predictions of equal dimensions to the input data that were aggregated into a single prediction vector prior to AUROC/AUPR performance measurement, rather than having 10 separate predictions per iteration each with AUROC and AUPR measurements. Collectively, this procedure left 10 AUROC and 10 AUPR values, one for each repeat of the 10-fold cross-validation. These ten values were used to estimate the 99% confidence intervals of performance and were overlaid on plots with the average performance and confidence interval ribbons (Figure 12A).

Regarding plotting, we adapted an approach from the scikit-learn python package (https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html) in R to estimate the average AUROC and AUPR curves among their 10 repeated iterations. This can be a challenging task because the specificity breaks of the ten model iterations are not always equivalent to each other, requiring interpolation. Specifically, to obtain the average performance lines, we performed linear interpolation using the *approx()* base R function of each ROC and PR curve across 1000 equally spaced points between 0 and 1, also ensuring that each average curve begins and ends at the corners of the plots. The 1000 interpolated y-values between x=0 and x=1 were then used to calculate

the average ROC and PR curve and its concomitant 99% confidence interval at each point. Overlaying these average performance lines with 99% confidence interval ribbons showed good concordance (Figure 12A).

Immunotherapy response predictions

A small number of patients with melanoma and lung cancer in the UCSD cohort had clinical immunotherapy response information available. Due to small sample sizes, machine learning on these patients was done using nested leave-one-out cross-validation, such that each k^{th} patient was iteratively left out and a model was built on the *k-1* patients (tuned using internal four-fold cross-validation) to make a prediction about the immunotherapy response of the k^{th} patient. After iterating through all *k* patients, the list of predicted responses and known responses were compared to calculate ROC and PR curves and their respective areas. Using WIS-overlapping fungi, moderate discrimination between responders and non-responders was observed in patients with melanoma (data not shown) but not in lung cancer (data not shown).

Scrambled and shuffled control analyses

In addition to comparing ML model performances to null AUROC and AUPR values, we wanted to implement additional negative control analyses. These were done in two independent ways just prior to ML model building: (1) scrambling metadata of prediction labels and (2) shuffling the sample IDs in the count data. We note that the scrambling and shuffling can occur globally (i.e., once before all ML models are built and tested) or dynamically (i.e., just prior to ML model building but after data subsetting and labelling). For example, when discriminating one cancer type versus all others, global scrambling would randomly sample all disease type labels among all sample types, whereas dynamic scrambling would happen only after subsetting to primary tumors and relabeling the disease types to two classes (i.e., the cancer type of interest and "Others"). We tested both of these approaches and found that both generally worked; however, the dynamic scrambling and shuffling yielded more consistent results (less variance) and showed greater agreement with known null values (i.e., 50% AUROC and positive class prevalence for AUPR). Hence, we used dynamic scrambling and shuffling as negative controls when comparing performance to actual samples.

Taxonomic generalizability

To test taxonomic generalizability, we aggregated raw read counts based on the decontaminated fungal data up the taxonomic levels (species through phylum) prior to ML using the phyloseq R package (v. 1.38.0) (*96*). Aggregated counts were then inputted into the same ten-fold cross-validation models (repeated once) described above to estimate performance and concomitant 95% confidence intervals (data not shown).

Stratified halves validation analyses

As another control, we split raw TCGA count data into two stratified halves using sequencing center, sample type, and disease type metadata information. We again note that experimental strategy was covered in this stratification since 7 of 8 sequencing centers only performed WGS or RNA-Seq, and the one that did both [Broad] processed 83% of samples with WGS only. We then used both of these stratified halves to iteratively train ML models employing ten-fold cross-validation (repeated once) predicting one cancer type versus all others; each trained model was then immediately applied to the data of the other stratified half to discriminate that particular cancer type. The ML performance from testing each model on the corresponding half was then compared, revealing highly concordant values (data not shown). This process was repeated using Voom-SNM normalization as well with the same procedure except that Voom-SNM normalization occurred independently on each half after stratification but prior to ML model building/testing. This additional analysis showed highly concordant performance among TCGA primary tumors (data not shown).

TCGA, Hopkins, and UCSD cohorts: Statistical analyses

Downstream analyses and plots were generated with either R version 4.03 or 4.1.1. Common R packages used include phyloseq (v. 1.38.0), vegan (v.2.5-7), microbiome (v. 1.16.0), doMC (1.3.7), dplyr (v. 1.0.7), reshape2 (v. 1.4.4), ggpubr (0.4.0), ggsci (v. 2.9), rstatix (v. 0.7.0), ggrepel (v. 0.9.1), tibble (3.1.6), caret (6.0-90), gbm (v. 2.1.8), xgboost (v. 1.5.0.1), MLmetrics (v. 1.1.1), PRROC (v. 1.3.1), e1071 (v. 1.7-9), gmodels (v. 2.18.1), ANCOMBC (v. 1.4.0), decontam (v. 1.14.0), limma (v. 3.50.0), edgeR (v. 3.36.0), snm (v. 1.42.0), biomformat (v. 1.22.0), and Rhdf5lib (v. 1.16.0). The rstatix package corrected for multiple hypothesis testing where applicable. Sample sizes were not estimated in advance and power calculations were not performed. The gbm package

was used for two-class gradient boosting ML and the xgboost package was used for multi-class gradient boosting ML. AUROC and AUPR were calculated using the PRROC package.

Appendices

Supplementary Tables (S1-S11)

Available at:
https://docs.google.com/spreadsheets/d/1qQIg3cEmyHCcyDPhRhE9K7w_7Apv4lgf/edit?usp=sharing&ouid=108555372942146520552&rtpof=true&sd=true

Table S1. Raw read counts of ASVs per sample in WIS cohort

Table S2. Floored, normalized read counts of ASVs per sample in WIS cohort

Table S3. Floored, normalized reads of all samples agglomerated to each taxonomic level. Table includes clinical data.

Table S4. Taxonomic classification and contamination filter status of ASVs in WIS cohort

Table S5. Detailed breakdown of sample cohorts

Table S6. In-depth contamination analysis for TCGA fungal calls

Table S7. Pan-TCGA fungal genome coverages

Table S8. WIS fungi-bacteria network analysis results with FDR less than 0.25

Table S9. PERMANOVA analysis within each disease type for sample-type for both Aitchison and Bray-Curtis distances to compare tumor vs. NAT

Table S10. WIS-overlapping fungal and bacterial genera used in TCGA MMvec co-occurrence analyses

Table S11. Top 20 fungi in Hopkins cohort discriminating pan-cancer vs. healthy in 10-fold CV repeated 10-times models

Bibliography

1. Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature*. **486**, 207–214 (2012).

2. J. Lloyd-Price, G. Abu-Ali, C. Huttenhower, The healthy human microbiome. *Genome Med.* **8**, 51 (2016).

3. S. V. Rajagopala, S. Vashee, L. M. Oldfield, Y. Suzuki, J. C. Venter, A. Telenti, K. E. Nelson, The Human Microbiome and Cancer. *Cancer Prev. Res.* . **10**, 226–234 (2017).

4. Y. Belkaid, T. W. Hand, Role of the microbiota in immunity and inflammation. *Cell*. **157**, 121–141 (2014).

5. J. A. Gilbert, R. A. Quinn, J. Debelius, Z. Z. Xu, J. Morton, N. Garg, J. K. Jansson, P. C. Dorrestein, R. Knight, Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*. **535**, 94–103 (2016).

6. R. F. Schwabe, C. Jobin, The microbiome and cancer. *Nat. Rev. Cancer*. **13**, 800–812 (2013).

7. Food Forum, Food and Nutrition Board, Institute of Medicine, *The Human Microbiome, Diet, and Health: Workshop Summary* (National Academies Press, 2013).

8. L. T. Geller, M. Barzily-Rokni, T. Danino, O. H. Jonas, N. Shental, D. Nejman, N. Gavert, Y. Zwang, Z. A. Cooper, K. Shee, C. A. Thaiss, A. Reuben, J. Livny, R. Avraham, D. T. Frederick, M. Ligorio, K. Chatman, S. E. Johnston, C. M. Mosher, A. Brandis, G. Fuks, C. Gurbatri, V. Gopalakrishnan, M. Kim, M. W. Hurd, M. Katz, J. Fleming, A. Maitra, D. A. Smith, M. Skalak, J. Bu, M. Michaud, S. A. Trauger, I. Barshack, T. Golan, J. Sandbank, K. T. Flaherty, A. Mandinova, W. S. Garrett, S. P. Thayer, C. R. Ferrone, C. Huttenhower, S. N. Bhatia, D. Gevers, J. A. Wargo, T. R. Golub, R. Straussman, Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*. **357**, 1156–1160 (2017).

9. S. Pushalkar, M. Hundeyin, D. Daley, C. P. Zambirinis, E. Kurz, A. Mishra, N. Mohan, B. Aykut, M. Usyk, L. E. Torres, G. Werba, K. Zhang, Y. Guo, Q. Li, N. Akkad, S. Lall, B. Wadowski, J. Gutierrez, J. A. Kochen Rossi, J. W. Herzog, B. Diskin, A. Torres-Hernandez, J. Leinwand, W. Wang, P. S. Taunk, S. Savadkar, M. Janal, A. Saxena, X. Li, D. Cohen, R. B. Sartor, D. Saxena, G. Miller, The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov.* **8**, 403–416 (2018).

10. S. Kalaora, A. Nagler, D. Nejman, M. Alon, C. Barbolin, E. Barnea, S. L. C. Ketelaars, K. Cheng, K. Vervier, N. Shental, Y. Bussi, R. Rotkopf, R. Levy, G. Benedek, S. Trabish, T. Dadosh, S. Levin-Zaidman, L. T. Geller, K. Wang, P. Greenberg, G. Yagel, A. Peri, G. Fuks, N. Bhardwaj, A. Reuben, L. Hermida, S. B. Johnson, J. R. Galloway-Peña, W. C. Shropshire, C. Bernatchez, C. Haymaker, R. Arora, L. Roitman, R. Eilam, A. Weinberger, M. Lotan-Pompan, M. Lotem, A. Admon, Y. Levin, T. D. Lawley, D. J. Adams, M. P. Levesque, M. J. Besser, J. Schachter, O. Golani, E. Segal, N. Geva-Zatorsky, E. Ruppin, P. Kvistborg, S. N. Peterson, J. A. Wargo, R. Straussman, Y. Samuels, Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature*. **592**, 138–143 (2021).

11. J.-C. J. Tsay, B. G. Wu, I. Sulaiman, K. Gershner, R. Schluger, Y. Li, T.-A. Yie, P. Meyn, E. Olsen, L. Perez, B. Franca, J. Carpenito, T. Iizumi, M. El-Ashmawy, M. Badri, J. T. Morton, N. Shen, L. He, G. Michaud, S. Rafeq, J. L. Bessich, R. L. Smith, H. Sauthoff, K. Felner, R. Pillai, A.-M. Zavitsanou, S. B. Koralov, V. Mezzano, C. A. Loomis, A. L. Moreira, W. Moore, A. Tsirigos, A. Heguy, W. N. Rom, D. H. Sterman, H. I. Pass, J. C. Clemente, H. Li, R. Bonneau, K.-K. Wong, T. Papagiannakopoulos, L. N. Segal, Lower Airway Dysbiosis Affects Lung Cancer Progression. *Cancer Discov.* **11**, 293–307 (2021).

12. E. Riquelme, Y. Zhang, L. Zhang, M. Montiel, M. Zoltan, W. Dong, P. Quesada, I. Sahin, V. Chandra, A. San Lucas, P. Scheet, H. Xu, S. M. Hanash, L. Feng, J. K. Burks, K.-A. Do, C. B. Peterson, D. Nejman, C.-W. D. Tzeng, M. P. Kim, C. L. Sears, N. Ajami, J. Petrosino, L. D. Wood, A. Maitra, R. Straussman, M. Katz, J. R. White, R. Jenq, J. Wargo, F. McAllister, Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell*. **178**, 795–806.e12 (2019).

13. L. Parhi, T. Alon-Maimon, A. Sol, D. Nejman, A. Shhadeh, T. Fainsod-Levi, O. Yajuk, B. Isaacson, J. Abed, N. Maalouf, A. Nissan, J. Sandbank, E. Yehuda-Shnaidman, F. Ponath, J. Vogel, O. Mandelboim, Z. Granot, R. Straussman, G. Bachrach, Breast cancer colonization by Fusobacterium nucleatum accelerates tumor growth and metastatic progression. *Nat. Commun.* **11**, 3259 (2020).

14. V. Le Noci, S. Guglielmetti, S. Arioli, C. Camisaschi, F. Bianchi, M. Sommariva, C. Storti, T. Triulzi, C. Castelli, A. Balsari, E. Tagliabue, L. Sfondrini, Modulation of Pulmonary Microbiota by Antibiotic or Probiotic Aerosol Therapy: A Strategy to Promote Immunosurveillance against Lung Metastases. *Cell Rep.* **24**, 3528–3538 (2018).

15. Y. Shi, W. Zheng, K. Yang, K. G. Harris, K. Ni, L. Xue, W. Lin, E. B. Chang, R. R. Weichselbaum, Y.-X. Fu, Intratumoral accumulation of gut microbiota facilitates CD47-based immunotherapy via STING signaling. *J. Exp. Med.* **217** (2020), doi:10.1084/jem.20192282.

16. K. Mima, Y. Sukawa, R. Nishihara, Z. R. Qian, M. Yamauchi, K. Inamura, S. A. Kim, A. Masuda, J. A. Nowak, K. Nosho, A. D. Kostic, M. Giannakis, H. Watanabe, S. Bullman, D. A. Milner, C. C. Harris, E. Giovannucci, L. A. Garraway, G. J. Freeman, G. Dranoff, A. T. Chan, W. S. Garrett, C. Huttenhower, C. S. Fuchs, S. Ogino, Fusobacterium nucleatum and T Cells in

Colorectal Carcinoma. *JAMA Oncol*. **1**, 653–661 (2015).

17. D. Nejman, I. Livyatan, G. Fuks, N. Gavert, Y. Zwang, L. T. Geller, A. Rotter-Maskowitz, R. Weiser, G. Mallel, E. Gigi, A. Meltser, G. M. Douglas, I. Kamer, V. Gopalakrishnan, T. Dadosh, S. Levin-Zaidman, S. Avnet, T. Atlan, Z. A. Cooper, R. Arora, A. P. Cogdill, M. A. W. Khan, G. Ologun, Y. Bussi, A. Weinberger, M. Lotan-Pompan, O. Golani, G. Perry, M. Rokah, K. Bahar-Shany, E. A. Rozeman, C. U. Blank, A. Ronai, R. Shaoul, A. Amit, T. Dorfman, R. Kremer, Z. R. Cohen, S. Harnof, T. Siegal, E. Yehuda-Shnaidman, E. N. Gal-Yam, H. Shapira, N. Baldini, M. G. I. Langille, A. Ben-Nun, B. Kaufman, A. Nissan, T. Golan, M. Dadiani, K. Levanon, J. Bar, S. Yust-Katz, I. Barshack, D. S. Peeper, D. J. Raz, E. Segal, J. A. Wargo, J. Sandbank, N. Shental, R. Straussman, The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science*. **368**, 973–980 (2020).

18. G. D. Poore, E. Kopylova, Q. Zhu, C. Carpenter, S. Fraraccio, S. Wandro, T. Kosciolek, S. Janssen, J. Metcalf, S. J. Song, J. Kanbar, S. Miller-Montgomery, R. Heaton, R. Mckay, S. P. Patel, A. D. Swafford, R. Knight, Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*. **579**, 567–574 (2020).

19. A. Wieland, M. R. Patel, M. A. Cardenas, C. S. Eberhardt, W. H. Hudson, R. C. Obeng, C. C. Griffith, X. Wang, Z. G. Chen, H. T. Kissick, N. F. Saba, R. Ahmed, Defining HPV-specific B cell responses in patients with head and neck cancer. *Nature*. **597**, 274–278 (2021).

20. M. Zapatka, I. Borozan, D. S. Brewer, M. Iskar, A. Grundhoff, M. Alawi, N. Desai, H. Sültmann, H. Moch, PCAWG Pathogens, C. S. Cooper, R. Eils, V. Ferretti, P. Lichter, PCAWG Consortium, The landscape of viral associations in human cancers. *Nat. Genet.* **52**, 320–330 (2020).

21. J. D. Forbes, C. N. Bernstein, H. Tremlett, G. Van Domselaar, N. C. Knox, A Fungal World: Could the Gut Mycobiome Be Involved in Neurological Disease? *Front. Microbiol.* **9**, 3249 (2018).

22. M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, MetaHIT Consortium, M. Antolín, F. Artiguenave, H. M. Blottiere, M. Almeida, C. Brechot, C. Cara, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, K. U. Foerstner, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, W. Huber, J. van Hylckama-Vlieg, A. Jamet, C. Juste, G. Kaci, J. Knol, O. Lakhdari, S. Layec, K. Le Roux, E. Maguin, A. Mérieux, R. Melo Minardi, C. M'rini, J. Muller, R. Oozeer, J. Parkhill, P. Renault, M. Rescigno, N. Sanchez, S. Sunagawa, A. Torrejon, K. Turner, G. Vandemeulebrouck, E. Varela, Y. Winogradsky, G. Zeller, J. Weissenbach, S. D. Ehrlich, P. Bork, Enterotypes of the human gut microbiome. *Nature*. **473**, 174–180 (2011).

23. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, MetaHIT Consortium, P. Bork, S. D. Ehrlich, J. Wang, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. **464**, 59–65 (2010).

24. I. D. Iliev, V. A. Funari, K. D. Taylor, Q. Nguyen, C. N. Reyes, S. P. Strom, J. Brown, C. A. Becker, P. R. Fleshner, M. Dubinsky, J. I. Rotter, H. L. Wang, D. P. B. McGovern, G. D. Brown, D. M. Underhill, Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. *Science*. **336**, 1314–1317 (2012).

25. K. S. Ost, T. R. O'Meara, W. Z. Stephens, T. Chiaro, H. Zhou, J. Penman, R. Bell, J. R. Catanzaro, D. Song, S. Singh, D. H. Call, E. Hwang-Wong, K. E. Hanson, J. F. Valentine, K. A. Christensen, R. M. O'Connell, B. Cormack, A. S. Ibrahim, N. W. Palm, S. M. Noble, J. L. Round, Adaptive immunity induces mutualism between commensal eukaryotes. *Nature*. **596**, 114–118 (2021).

26. U. Jain, A. M. Ver Heul, S. Xiong, M. H. Gregory, E. G. Demers, J. T. Kern, C.-W. Lai, B. D. Muegge, D. A. G. Barisas, J. S. Leal-Ekman, P. Deepak, M. A. Ciorba, T.-C. Liu, D. A. Hogan, P. Debbas, J. Braun, D. P. B. McGovern, D. M. Underhill, T. S. Stappenbeck, Debaryomyces is enriched in Crohn's disease intestinal tissue and impairs healing in mice. *Science*. **371**, 1154–1159 (2021).

27. J. R. Köhler, A. Casadevall, J. Perfect, The spectrum of fungi that infects humans. *Cold Spring Harb. Perspect. Med.* **5**, a019273 (2014).

28. J. R. Galloway-Peña, D. P. Kontoyiannis, The gut mycobiome: The overlooked constituent of clinical outcomes and treatment complications in patients with cancer and other immunosuppressive conditions. *PLoS Pathog.* **16**, e1008353 (2020).

29. J. J. Limon, J. H. Skalski, D. M. Underhill, Commensal Fungi in Health and Disease. *Cell Host Microbe*. **22**, 156–165 (2017).

30. H. E. Hallen-Adams, S. D. Kachman, J. Kim, R. M. Legge, I. Martínez, Fungi inhabiting the healthy human gastrointestinal tract: a diverse and dynamic community. *Fungal Ecol.* **15**, 9–17 (2015).

31. T. Heisel, E. Montassier, A. Johnson, G. Al-Ghalith, Y.-W. Lin, L.-N. Wei, D. Knights, C. A. Gale, High-Fat Diet Changes Fungal Microbiomes and Interkingdom Relationships in the Murine Gut. *mSphere*. **2** (2017), doi:10.1128/mSphere.00351-17.

32. C. E. Huseyin, P. W. O'Toole, P. D. Cotter, P. D. Scanlan, Forgotten fungi—the gut mycobiome in human health and disease. *FEMS Microbiol. Rev.* **41**, 479–511 (2017).

33. H. Sokol, V. Leducq, H. Aschard, H.-P. Pham, S. Jegou, C. Landman, D. Cohen, G. Liguori, A. Bourrier, I. Nion-Larmurier, J. Cosnes, P. Seksik, P. Langella, D. Skurnik, M. L. Richard, L. Beaugerie, Fungal microbiota dysbiosis in IBD. *Gut*. **66**, 1039–1048 (2017).

34. K. Findley, J. Oh, J. Yang, S. Conlan, C. Deming, J. A. Meyer, D. Schoenfeld, E. Nomicos, M. Park, NIH Intramural Sequencing Center Comparative Sequencing Program, H. H. Kong, J. A. Segre, Topographic diversity of fungal and bacterial communities in human skin. *Nature*. **498**, 367–370 (2013).

35. J.-H. Jo, C. Deming, E. A. Kennedy, S. Conlan, E. C. Polley, W.-I. Ng, NISC Comparative Sequencing Program, J. A. Segre, H. H. Kong, Diverse Human Skin Fungal Communities in Children Converge in Adulthood. *J. Invest. Dermatol.* **136**, 2356–2363 (2016).

36. A. K. Dupuy, M. S. David, L. Li, T. N. Heider, J. D. Peterson, E. A. Montano, A. Dongari-Bagtzoglou, P. I. Diaz, L. D. Strausbaugh, Redefining the Human Oral Mycobiome with Improved Practices in Amplicon-based Taxonomy: Discovery of Malassezia as a Prominent Commensal. *PLoS ONE*. **9** (2014), p. e90899.

37. M. A. Ghannoum, R. J. Jurevic, P. K. Mukherjee, F. Cui, M. Sikaroodi, A. Naqvi, P. M. Gillevet, Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6**, e1000713 (2010).

38. P. K. Mukherjee, J. Chandra, M. Retuerto, M. Sikaroodi, R. E. Brown, R. Jurevic, R. A. Salata, M. M. Lederman, P. M. Gillevet, M. A. Ghannoum, Oral mycobiome analysis of HIV-infected patients: identification of Pichia as an antagonist of opportunistic fungi. *PLoS Pathog.* **10**, e1003996 (2014).

39. M. B. Lawani, A. Morris, The respiratory microbiome of HIV-infected individuals. *Expert Rev. Anti. Infect. Ther.* **14**, 719–729 (2016).

40. L. D. N. Nguyen, E. Viscogliosi, L. Delhaes, The lung mycobiome: an emerging field of the human respiratory microbiome. *Frontiers in Microbiology*. **6** (2015), , doi:10.3389/fmicb.2015.00089.

41. I. D. Iliev, I. Leonardi, Fungal dysbiosis: immunity and interactions at mucosal barriers. *Nat. Rev. Immunol.* **17**, 635–646 (2017).

42. O. O. Coker, G. Nakatsu, R. Z. Dai, W. K. K. Wu, S. H. Wong, S. C. Ng, F. K. L. Chan, J. J. Y. Sung, J. Yu, Enteric fungal microbiota dysbiosis and ecological alterations in colorectal cancer. *Gut*. **68** (2019), pp. 654–662.

43. R. Gao, C. Kong, H. Li, L. Huang, X. Qu, N. Qin, H. Qin, Dysbiosis signature of mycobiota in colon polyp and colorectal cancer. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 2457–2468 (2017).

44. M. J. Paterson, S. Oh, D. M. Underhill, Host–microbe interactions: commensal fungi in the gut. *Current Opinion in Microbiology*. **40** (2017), pp. 131–137.

45. S. Banerjee, Z. Wei, F. Tan, K. N. Peck, N. Shih, M. Feldman, T. R. Rebbeck, J. C. Alwine, E. S. Robertson, Distinct microbiological signatures associated with

triple negative breast cancer. *Sci. Rep.* **5**, 15162 (2015).

46. S. Banerjee, T. Tian, Z. Wei, N. Shih, M. D. Feldman, K. N. Peck, A. M. DeMichele, J. C. Alwine, E. S. Robertson, Distinct Microbial Signatures Associated With Different Breast Cancer Types. *Front. Microbiol.* **9**, 951 (2018).

47. S. Banerjee, T. Tian, Z. Wei, N. Shih, M. D. Feldman, J. C. Alwine, G. Coukos, E. S. Robertson, The ovarian cancer oncobiome. *Oncotarget*. **8**, 36225–36245 (2017).

48. S. Banerjee, J. C. Alwine, Z. Wei, T. Tian, N. Shih, C. Sperling, T. Guzzo, M. D. Feldman, E. S. Robertson, Microbiome signatures in prostate cancer. *Carcinogenesis*. **40**, 749–764 (2019).

49. B. Aykut, S. Pushalkar, R. Chen, Q. Li, R. Abengozar, J. I. Kim, S. A. Shadaloey, D. Wu, P. Preiss, N. Verma, Y. Guo, A. Saxena, M. Vardhan, B. Diskin, W. Wang, J. Leinwand, E. Kurz, J. A. Kochen Rossi, M. Hundeyin, C. Zambrinis, X. Li, D. Saxena, G. Miller, The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature*. **574**, 264–267 (2019).

50. F. Zhu, J. Willette-Brown, N.-Y. Song, D. Lomada, Y. Song, L. Xue, Z. Gray, Z. Zhao, S. R. Davis, Z. Sun, P. Zhang, X. Wu, Q. Zhan, E. R. Richie, Y. Hu, Autoreactive T Cells and Chronic Fungal Infection Drive Esophageal Carcinogenesis. *Cell Host Microbe*. **21**, 478–493.e7 (2017).

51. A. Ramirez-Garcia, A. Rementeria, J. M. Aguirre-Urizar, M. D. Moragues, A. Antoran, A. Pellon, A. Abad-Diaz-de-Cerio, F. L. Hernando, Candida albicansand cancer: Can this yeast induce cancer development or progression? *Critical Reviews in Microbiology* (2014), pp. 1–13.

52. Y. Masuda, H. Inoue, H. Ohta, A. Miyake, M. Konishi, H. Nanba, Oral administration of soluble β-glucans extracted fromGrifola frondosainduces systemic antitumor immune response and decreases immunosuppression in tumor-bearing mice. *International Journal of Cancer*. **133** (2013), pp. 108–119.

53. Y. Masuda, Y. Nakayama, A. Tanaka, K. Naito, M. Konishi, Antitumor activity of orally administered maitake α-glucan by stimulating antitumor immune response in murine tumor. *PLOS ONE*. **12** (2017), p. e0173621.

54. A. Y. Peleg, D. A. Hogan, E. Mylonakis, Medically important bacterial–fungal interactions. *Nature Reviews Microbiology*. **8** (2010), pp. 340–349.

55. S. L. Shiao, K. M. Kershaw, J. J. Limon, S. You, J. Yoon, E. Y. Ko, J. Guarnerio, A. A. Potdar, D. P. B. McGovern, S. Bose, T. B. Dar, P. Noe, J. Lee, Y. Kubota, V. I. Maymi, M. J. Davis, R. M. Henson, R. Y. Choi, W. Yang, J. Tang, M. Gargus, A. D. Prince, Z. S. Zumsteg, D. M. Underhill, Commensal bacteria and fungi differentially regulate tumor responses to radiation therapy. *Cancer Cell*. **39** (2021), pp. 1202–1213.e6.

56. P. Frey-Klett, P. Burlinson, A. Deveau, M. Barret, M. Tarkka, A. Sarniguet, Bacterial-Fungal Interactions: Hyphens between Agricultural, Clinical, Environmental, and Food Microbiologists. *Microbiology and Molecular Biology*

Reviews. **75** (2011), pp. 583–609.

57. J. Inácio, S. Behrens, B. M. Fuchs, A. Fonseca, I. Spencer-Martins, R. Amann, In situ accessibility of Saccharomyces cerevisiae 26S rRNA to Cy3-labeled oligonucleotide probes comprising the D1 and D2 domains. *Appl. Environ. Microbiol.* **69**, 2899–2905 (2003).

58. A. K. Nash, T. A. Auchtung, M. C. Wong, D. P. Smith, J. R. Gesell, M. C. Ross, C. J. Stewart, G. A. Metcalf, D. M. Muzny, R. A. Gibbs, N. J. Ajami, J. F. Petrosino, The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome*. **5** (2017), , doi:10.1186/s40168-017-0373-4.

59. S. Reitmeier, T. C. A. Hitch, N. Treichel, N. Fikas, B. Hausmann, A. E. Ramer-Tait, K. Neuhaus, D. Berry, D. Haller, I. Lagkouvardos, T. Clavel, Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications*. **1** (2021), , doi:10.1038/s43705-021-00033-z.

60. S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. Ø. Jensen, J. E. Medina, C. Hruban, J. R. White, D. N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B. D. Woodward, H. Husain, K. L. van Rooijen, M.-B. W. Ørntoft, A. H. Madsen, C. J. H. van de Velde, M. Verheij, A. Cats, C. J. A. Punt, G. R. Vink, N. C. T. van Grieken, M. Koopman, R. J. A. Fijneman, J. S. Johansen, H. J. Nielsen, G. A. Meijer, C. L. Andersen, R. B. Scharpf, V. E. Velculescu, Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. **570**, 385–389 (2019).

61. G. D. Sepich-Poore, L. Zitvogel, R. Straussman, J. Hasty, J. A. Wargo, R. Knight, The microbiome and human cancer. *Science*. **371** (2021), , doi:10.1126/science.abc4552.

62. S. Banerjee, Z. Wei, T. Tian, D. Bose, N. N. C. Shih, M. D. Feldman, T. Khoury, A. De Michele, E. S. Robertson, Prognostic correlations with the microbiome of breast cancer subtypes. *Cell Death & Disease*. **12** (2021), , doi:10.1038/s41419-021-04092-x.

63. N. Howlader, S. F. Altekruse, C. I. Li, V. W. Chen, C. A. Clarke, L. A. G. Ries, K. A. Cronin, US incidence of breast cancer subtypes defined by joint hormone receptor and HER2 status. *J. Natl. Cancer Inst.* **106** (2014), doi:10.1093/jnci/dju055.

64. A. Boix-Amorós, C. Martinez-Costa, A. Querol, M. C. Collado, A. Mira, Multiple Approaches Detect the Presence of Fungi in Human Breastmilk Samples from Healthy Mothers. *Sci. Rep.* **7**, 13016 (2017).

65. T. M. Neeson, Y. Mandelik, Pairwise measures of species co-occurrence for choosing indicator species and quantifying overlap. *Ecological Indicators*. **45** (2014), pp. 721–727.

66. D. M. Underhill, I. D. Iliev, The mycobiota: interactions between commensal fungi and the host immune system. *Nat. Rev. Immunol.* **14**, 405–416 (2014).

67. A. J. Wolf, D. M. Underhill, Peptidoglycan recognition by the innate immune system. *Nature Reviews Immunology*. **18** (2018), pp. 243–254.

68. V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, D. S. Bortone, T.-H. Ou Yang, E. Porta-Pardo, G. F. Gao, C. L. Plaisier, J. A. Eddy, E. Ziv, A. C. Culhane, E. O. Paull, I. K. A. Sivakumar, A. J. Gentles, R. Malhotra, F. Farshidfar, A. Colaprico, J. S. Parker, L. E. Mose, N. S. Vo, J. Liu, Y. Liu, J. Rader, V. Dhankani, S. M. Reynolds, R. Bowlby, A. Califano, A. D. Cherniack, D. Anastassiou, D. Bedognetti, Y. Mokrab, A. M. Newman, A. Rao, K. Chen, A. Krasnitz, H. Hu, T. M. Malta, H. Noushmehr, C. S. Pedamallu, S. Bullman, A. I. Ojesina, A. Lamb, W. Zhou, H. Shen, T. K. Choueiri, J. N. Weinstein, J. Guinney, J. Saltz, R. A. Holt, C. S. Rabkin, Cancer Genome Atlas Research Network, A. J. Lazar, J. S. Serody, E. G. Demicco, M. L. Disis, B. G. Vincent, I. Shmulevich, The Immune Landscape of Cancer. *Immunity*. **48**, 812–830.e14 (2018).

69. A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, A. A. Alizadeh, Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*. **12** (2015), pp. 453–457.

70. J. T. Morton, A. A. Aksenov, L. F. Nothias, J. R. Foulds, R. A. Quinn, M. H. Badri, T. L. Swenson, M. W. Van Goethem, T. R. Northen, Y. Vazquez-Baeza, M. Wang, N. A. Bokulich, A. Watters, S. J. Song, R. Bonneau, P. C. Dorrestein, R. Knight, Learning representations of microbe-metabolite interactions. *Nat. Methods*. **16**, 1306–1314 (2019).

71. C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List, Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6241–6246 (2012).

72. A. L. Torres-Machorro, R. Hernández, A. M. Cevallos, I. López-Villaseñor, Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny? *FEMS Microbiology Reviews*. **34** (2010), pp. 59–86.

73. J. Tang, I. D. Iliev, J. Brown, D. M. Underhill, V. A. Funari, Mycobiome: Approaches to analysis of intestinal fungi. *J. Immunol. Methods*. **421**, 112–121 (2015).

74. T. Kobayashi, D. J. Heck, M. Nomura, T. Horiuchi, Expansion and contraction of ribosomal DNA repeats in Saccharomyces cerevisiae: requirement of replication fork blocking (Fob1) protein and the role of RNA polymerase I. *Genes Dev.* **12**, 3821–3830 (1998).

75. M. Op De Beeck, B. Lievens, P. Busschaert, S. Declerck, J. Vangronsveld, J. V. Colpaert, Comparison and validation of some ITS primer pairs useful for fungal metabarcoding studies. *PLoS One*. **9**, e97629 (2014).

76. C. Luan, L. Xie, X. Yang, H. Miao, N. Lv, R. Zhang, X. Xiao, Y. Hu, Y. Liu, N. Wu, Y. Zhu, B. Zhu, Dysbiosis of fungal microbiota in the intestinal mucosa of

patients with colorectal adenomas. *Sci. Rep.* **5**, 7980 (2015).

77. M. G. I. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, C. Huttenhower, Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

78. J. M. Bain, M. F. Alonso, D. S. Childers, C. A. Walls, K. Mackenzie, A. Pradhan, L. E. Lewis, J. Louw, G. M. Avelar, D. E. Larcombe, M. G. Netea, N. A. R. Gow, G. D. Brown, L. P. Erwig, A. J. P. Brown, Immune cells fold and damage fungal hyphae. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021), doi:10.1073/pnas.2020484118.

79. A. S. Gilbert, R. T. Wheeler, R. C. May, Fungal Pathogens: Survival and Replication within Macrophages. *Cold Spring Harbor Perspectives in Medicine*. **5** (2015), p. a019661.

80. A. Frau, J. G. Kenny, L. Lenzi, B. J. Campbell, U. Z. Ijaz, C. A. Duckworth, M. D. Burkitt, N. Hall, J. Anson, A. C. Darby, C. S. J. Probert, DNA extraction and amplicon production strategies deeply inf luence the outcome of gut mycobiome studies. *Sci. Rep.* **9**, 9328 (2019).

81. R. Tignat-Perrier, A. Dommergue, A. Thollot, C. Keuschnig, O. Magand, T. M. Vogel, C. Larose, Global airborne microbial communities controlled by surrounding landscapes and wind conditions. *Sci. Rep.* **9**, 14441 (2019).

82. G. C.-F. Chan, W. K. Chan, D. M.-Y. Sze, The effects of β-glucan on human immune and cancer cells. *J. Hematol. Oncol.* **2**, 1–11 (2009).

83. B. Seelbinder, J. Chen, S. Brunke, R. Vazquez-Uribe, R. Santhaman, A.-C. Meyer, F. S. de Oliveira Lino, K.-F. Chan, D. Loos, L. Imamovic, C.-C. Tsang, R. P.-K. Lam, S. Sridhar, K. Kang, B. Hube, P. C.-Y. Woo, M. O. A. Sommer, G. Panagiotou, Antibiotics create a shift from mutualism to competition in human gut communities with a longer-lasting impact on fungi than bacteria. *Microbiome*. **8**, 133 (2020).

84. D. C. Hinshaw, L. A. Shevde, The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Research*. **79** (2019), pp. 4557–4566.

85. C. Wagg, K. Schlaeppi, S. Banerjee, E. E. Kuramae, M. G. A. van der Heijden, Fungal-bacterial diversity and microbiome complexity predict ecosystem functioning. *Nature Communications*. **10** (2019), , doi:10.1038/s41467-019-12798-y.

86. J. W. Lau, E. Lehnert, A. Sethi, R. Malhotra, G. Kaushik, Z. Onder, N. Groves-Kirkby, A. Mihajlovic, J. DiGiovanna, M. Srdic, D. Bajcic, J. Radenkovic, V. Mladenovic, D. Krstanovic, V. Arsenijevic, D. Klisic, M. Mitrovic, I. Bogicevic, D. Kural, B. Davis-Dusenbery, The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. *Cancer Res.* **77**, e3–e6 (2017).

87. K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, A.

M. Taylor, A. D. Cherniack, V. Thorsson, R. Akbani, R. Bowlby, C. K. Wong, M. Wiznerowicz, F. Sanchez-Vega, A. G. Robertson, B. G. Schneider, M. S. Lawrence, H. Noushmehr, T. M. Malta, Cancer Genome Atlas Network, J. M. Stuart, C. C. Benz, P. W. Laird, Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*. **173**, 291–304.e6 (2018).

88. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

89. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature*. **578**, 82–93 (2020).

90. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. **34**, 3094–3100 (2018).

91. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).

92. A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, J. G. Sanders, J. Shorenstein, H. Holste, S. Petrus, A. Robbins-Pianka, C. J. Brislawn, M. Wang, J. R. Rideout, E. Bolyen, M. Dillon, J. G. Caporaso, P. C. Dorrestein, R. Knight, Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*. **15**, 796–798 (2018).

93. Q. Zhu, S. Huang, A. Gonzalez, I. McGrath, D. McDonald, OGUs enable effective, phylogeny-aware analysis of even shallow metagenome community structures. *bioRxiv* (2021) (available at https://www.biorxiv.org/content/10.1101/2021.04.04.438427v1.abstract).

94. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

95. E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J. Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, R. Da Silva, C. Diener, P. C. Dorrestein, G. M. Douglas, D. M. Durall, C. Duvallet, C. F. Edwardson, M. Ernst, M. Estaki, J. Fouquier, J. M. Gauglitz, S. M. Gibbons, D. L. Gibson, A. Gonzalez, K. Gorlick, J. Guo, B. Hillmann, S. Holmes, H. Holste, C. Huttenhower, G. A. Huttley, S. Janssen, A. K. Jarmusch, L. Jiang, B. D. Kaehler, K. B. Kang, C. R. Keefe, P. Keim, S. T. Kelley, D. Knights, I. Koester, T. Kosciolek, J. Kreps, M. G. I. Langille, J. Lee, R. Ley, Y.-X. Liu, E. Loftfield, C. Lozupone, M. Maher, C. Marotz, B. D. Martin, D. McDonald, L. J. McIver, A. V. Melnik, J. L. Metcalf, S. C. Morgan, J. T. Morton, A. T. Naimey, J. A. Navas-Molina, L. F. Nothias, S. B. Orchanian, T. Pearson, S. L. Peoples, D. Petras, M. L. Preuss, E. Pruesse, L. B. Rasmussen, A. Rivers, M. S. Robeson 2nd, P. Rosenthal, N. Segata, M. Shaffer, A. Shiffer, R. Sinha, S. J. Song, J. R. Spear, A.

D. Swafford, L. R. Thompson, P. J. Torres, P. Trinh, A. Tripathi, P. J. Turnbaugh, S. Ul-Hasan, J. J. J. van der Hooft, F. Vargas, Y. Vázquez-Baeza, E. Vogtmann, M. von Hippel, W. Walters, Y. Wan, M. Wang, J. Warren, K. C. Weber, C. H. D. Williamson, A. D. Willis, Z. Z. Xu, J. R. Zaneveld, Y. Zhang, Q. Zhu, R. Knight, J. G. Caporaso, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

96. P. J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. **8**, e61217 (2013).

97. C. Martino, J. T. Morton, C. A. Marotz, L. R. Thompson, A. Tripathi, R. Knight, K. Zengler, A novel sparse compositional technique reveals microbial perturbations. mSystems 4: e00016-19 (2019).

98. Y. Vázquez-Baeza, M. Pirrung, A. Gonzalez, R. Knight, EMPeror: a tool for visualizing high-throughput microbial community data. *Gigascience*. **2**, 16 (2013).

99. N. M. Davis, D. M. Proctor, S. P. Holmes, D. A. Relman, B. J. Callahan, Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*. **6**, 226 (2018).

100. C. Allaband, A. Lingaraju, C. Martino, B. Russell, A. Tripathi, O. Poulsen, A. C. Dantas Machado, D. Zhou, J. Xue, E. Elijah, A. Malhotra, P. C. Dorrestein, R. Knight, G. G. Haddad, A. Zarrinpar, Intermittent Hypoxia and Hypercapnia Alter Diurnal Rhythms of Luminal Gut Microbiome and Metabolome. *mSystems*, e0011621 (2021).

101. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17**, 261–272 (2020).

102. Hunter, Matplotlib: A 2D Graphics Environment. **9**, 90–95 (2007).

103. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

104. B. H. Mecham, P. S. Nelson, J. D. Storey, Supervised normalization of microarrays. *Bioinformatics*. **26**, 1308–1315 (2010).

105. A. Scherer, *Batch Effects and Noise in Microarray Experiments: Sources and Solutions* (John Wiley & Sons, 2009).