RESEARCH ARTICLE

A comprehensive analysis of breast cancer microbiota and host gene expression

Kevin J. Thompson[1,2], James N. Ingle[3], Xiaojia Tang[1,2], Nicholas Chia[2,4], Patricio R. Jeraldo[2,4], Marina R. Walther-Antonio[2,4], Karunya K. Kandimalla[5], Stephen Johnson[1,2], Janet Z. Yao[2], Sean C. Harrington[2], Vera J. Suman[1], Liewei Wang[6], Richard L. Weinshilboum[6], Judy C. Boughey[4], Jean-Pierre Kocher[1,2], Heidi Nelson[4], Matthew P. Goetz[3], Krishna R. Kalari[1,2] *

1 Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, United States of America, 2 Department of Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, United States of America, 3 Department of Oncology, Mayo Clinic, Rochester, Minnesota, United States of America, 4 Department of Surgery, Mayo Clinic, Rochester, Minnesota, United States of America, 5 Department of Pharmaceutics, University of Minnesota, Minneapolis, MN, United States of America, 6 Department of Molecular Pharmacology & Experimental Therapeutics, Mayo Clinic, Rochester, Minnesota, United States of America

* Kalari.Krishna@mayo.edu

Abstract

The inflammatory tumoral-immune response alters the physiology of the tumor microenvironment, which may attenuate genomic instability. In addition to inducing inflammatory immune responses, several pathogenic bacteria produce genotoxins. However the extent of microbial contribution to the tumor microenvironment biology remains unknown. We utilized The Cancer Genome Atlas, (TCGA) breast cancer data to perform a novel experiment utilizing unmapped and mapped RNA sequencing read evidence to minimize laboratory costs and effort. Our objective was to characterize the microbiota and associate the microbiota with the tumor expression profiles, for 668 breast tumor tissues and 72 non-cancerous adjacent tissues. The prominent presence of *Proteobacteria* was increased in the tumor tissues and conversely *Actinobacteria* abundance increase in non-cancerous adjacent tissues. Further, geneset enrichment suggests *Listeria spp* to be associated with the expression profiles of genes involved with epithelial to mesenchymal transitions. Moreover, evidence suggests *H. influenza* may reside in the surrounding stromal material and was significantly associated with the proliferative pathways: G2M checkpoint, E2F transcription factors, and mitotic spindle assembly. In summary, further unraveling this complicated interplay should enable us to better diagnose and treat breast cancer patients.

Introduction

Cancer is a complex disease, where a multitude of genomic and physiological alterations occurring incessantly in the tumor tissue adds to the complexity of disease treatment and management[1]. The microenvironment in and around the tumor constitutes a variety of cell types, purportedly including microbiota. Pathophysiological changes occurring in the cells and in the microbial composition could have significant impact on the tumor growth [1–4].

While impactful discoveries have been made in cancer diagnosis and treatment by investigating the intricate shifts in the cellular and molecular biology of the tumor, microbial contributions to tumor growth remain unexplored. Diversity shifts in tumor microbiota have been observed in a variety of cancers including: prostate cancer [5], cervical cancer [6], colorectal cancer [7, 8], and lung cancer [6, 9], and breast cancer [10–12].

Breast cancer is one of the most common cancers in women worldwide and despite the significant progress in the diagnosis and treatment of breast cancers, there are still more than 40,000 deaths per year [13, 14]. Dysregulation of sex hormones [13–15] is believed to be one of the primary risk factors for breast cancer. A menopause associated decrease in age-specific incidence rates, known as Clemmesen's hook, is widely observed among females worldwide [16, 17]. The hormonal dysregulation manifests, both clinically and molecularly, as distinct subtypes: triple negative (TN), HER2 positive (HER2+), and ER positive (ER+) [18]. Recently, it has been demonstrated that post-menopausal estrogen metabolism is associated with microbial diversity [19]. Similarly it has been proposed that estrogen conjugation by beta-glucuronidase, may be associated with the microbiota observed to be in dysbiosis in women with a history of breast cancer [20].

The objective of the current study is to characterize the breast microbiota and investigate whether the microbial composition is associated with host expression profiles. This objective was accomplished by utilizing RNA sequencing data from The Cancer Genome Atlas (TCGA), the largest sequencing cohort currently available for breast cancer [18]. We documented the most prevalent species observed among the breast tissues, and gathered preliminary evidence of microbial compositional shifts among the disease subtypes. This was further supported, by 16S-rRNA gene sequencing data obtained from fresh frozen samples that were obtained from the TCGA subjects submitted by the Mayo Clinic. Traditionally, bioinformatic methods such as PICRUSt are implemented to infer the abundance of gene families, which may in turn influence host's expression response [21]. In contrast, we utilized two lines of evidence- microbial read evidence and host transcriptional expression—from the same tissue and sequencing platform, to perform association analysis. The implementation of a single sequencing platform reduces confounding factors of tissue heterogeneity, molecule extraction/purification, and protocol handling differences. We believe this is the first study to examine both the microbial presence and host expression from the same tissue/sample preparation.

Materials and methods

RNA sequencing data

Clinical filtering. Clinical characteristics of the breast cancer samples were obtained from the TCGA data portal [22] (November 11, 2013, version). Samples were removed from consideration if they possessed any of the following characteristics: male gender, metastatic samples, history of prior breast cancer disease, recipients of neoadjuvant therapy, and/or samples lacking (documented) interrogation of HER2 amplification status. The fastq sequence files for the remaining 804 breast tumor and non-cancerous breast tumor, were obtained from the Cancer Genomics Hub (CGHuB) data repository[23].

Host aligned reads. We aligned RNA-Seq fastq files using TopHat (v1.3.3), mapped reads were used to obtain gene expression counts using HTSeq (v0.5.3p3) [24, 25]. Host gene expression counts were normalized with conditional quantile normalization to account for potential GC, and/or gene length biases [26]. Subtype specific host expression cohorts were inspected for outliers using calibration stress measures [27] and principal component analysis.

Non-host aligned reads. To characterize the breast microbiota, the reads unaligned to the host genome were aligned by Kraken to its bacterial genome (03–2014 build) and bacterial

16S ribosomal databases (06–2014 build) [28]. Subsequent to TopHat alignment unmapped BAM files, of the 804 tumor and non-cancerous adjacent (NCA) samples, were converted to FASTQ reads that were quality-trimmed with fastq_quality_trimmer (FASTX Toolkit v0.0.13.2). Read lengths less than 40 nucleotides and a minimum PHRED quality score of less than 3 were removed and the remainder were realigned with Bowtie 2 to eliminate any additional human reads. The left over unmapped reads were then aligned to bacterial genomes and bacterial 16S ribosomal genes using Kraken (v0.10.5-alpha) [28]. Concordance of sample retention, of the parallel processed TCGA samples, further pruned the study cohort to a final cohort consisting of 668 tumor samples and 72 NCA (NCA) reference samples.

16S ribosomal sequencing validation

Frozen breast tissue blocks demonstrating greater than 60% tumor composition were selected from six ER+ patients from Mayo Clinic that were submitted to TCGA. Frozen sections (15–20 at 4 µm) were cut from each block and stored at -80˚C prior to DNA extraction. The MoBio PowerSoil® DNA Isolation Kit (PN 12888 Mo Bio Laboratories, Inc. Carlsbad, CA) were used for DNA extraction according to the manufacturer's protocol. The DNA concentrations were measured by Qubit dsDNA HS Assay Kit (PN Q32854 Thermo Fisher Scientific Inc., Waltham, MA) and samples with sufficient DNA were enriched for microbial DNA using the NEBNext® Microbiome DNA Enrichment Kit (PN E2612L, New England Biolabs, Ipswich, MA). The V3-V5 region of the 16S-rRNA gene were amplified with a two-step PCR protocol, and then Illumina flow cell adaptors containing indices were incorporated [29]. Sequence reads (both pairs) were filtered for quality using PHRED quality scores Q3, and 3' reads with average sliding window (size = 4) score of Q15, using Trimmomatic v 0.22 [30]. Paired reads with that were at least 80% of the original read length nucleotides were aligned with the IM-TORNADO 16S analysis pipeline[31, 32] using the Greengenes taxonomy (Greengenes99 database version 12.10) [33]. Additional details are provided in S1 File.

Statistical analysis and visualization

Batch analysis. To account for dispersion and sparsity, microbial reads were normalized with the metagenomeSeq package [34]. We used the Bray-Curtis dissimilarity measure, a count-based measure to compare between two sampling sites, traditionally employed in ecological studies [35–37]. To identify the presence of handling/processing differences, centroid clustering was performed using the t-SNE dimensionality reduction on the Bray-Curtis dissimilarities [36, 38]. The NbClust package was then employed to determine the optimal number of clusters [39]. This package evaluated the validity of k (2:10) clusters, as scored by 26 indices. The 26 (default) of 30 potential indices were chosen as they are the least computationally expensive. Cluster validity was gaged by the average silhouette width, where a higher average silhouette width indicates high sample similarity (tightness) and substantial cluster separation[40, 41]. Clustering results were confirmed using DESeq2 normalized data and non-negative matrix factorization, see S2 File.

Statistical analysis. Concordances of 327 species were observed among the NCA tissues and the tumor samples, and consistent across the observed batch effect. The batch influence was adjusted for with the sva package, using the zero-inflated Gaussian normalized data [34, 42]. Differential abundance analysis of adjusted microbial read evidence was performed with limma for the most abundant operational taxonomic units (OTUs) [43]. Host differential gene expression analysis (human) was performed with edgeR, normalized to a negative binomial distribution to account for read dispersion [44]. Spearman's correlation was performed at the univariate level, among the OTU's and genes. Fisher's exact test was performed for gene set

enrichment analysis, using the 50 hallmark pathways [45, 46]. Additionally, mutual information was assessed, with the BUS package [47], among the normalized microbiome data and the 50 hallmark pathways.

Visualization. Bean plots were prepared with the Beanplot R package [48] and hive plots with HiveR [49]. Cladograms were produced with GraPhlAn software [50]. Parallel coordinates and parallel sets plots were generated with epade [51].

Results

Microbiota identification

The unaligned human genome reads obtained from 668 tumor and 72 NCA tissues were attributed to 1,634 microbial OTUs using dual alignment strategy, which was chosen to minimize contaminants [12, 52, 53]. We further applied a set of filters to remove rare microbial OTUs, those presented in less than 25% of our study cohort, and low abundant microbial OTUs with less than nine maximum reads across the samples. The elimination of OTUs occurring in less than 25% of the samples had little effect on the overall OTU counts of each sample, with 99.2% (657 samples) removing less than 5% of their overall read counts. An ER positive sample (TCGA-BH-A0BR) demonstrated the highest eliminated read counts (10.94%), see S1 Table. As microbial OTUs could be tissue specific, the tumor and NCA samples were grouped and processed in parallel. This left 662 OTUs observed among the tumor samples; 639 OTUs observed among the NCA tissues and with 589 OTUs observed concordantly in both tumor and NCA tissues (S1 Fig). We performed Kmeans clustering analysis using the reduced dimensions from t-distributed stochastic neighbor embedding (t-SNE). The t-SNE mappings were derived from the Bray-Curtis dissimilarity index. We observed that both tumor and NCA cohorts demonstrated the existence of a batch processing effect indicated by the TCGA plate id [54](S1 Fig).

Two clusters were identified for tumor (cluster 1 = 197 and cluster 2 = 471 samples), and NCA cohorts (cluster 1 = 54 and cluster 2 = 18 samples), Fig 1A. A PCA plot is provided in Fig 1B, demonstrating the sample separation created by the batch effect. The filtering describe above was applied to each of the clusters. The tumor cluster 1 (25%) and NCA cluster 2 (25%) demonstrated a concordant set of 198 OTU's (S2 Table) suggesting the same processing factor was influencing the samples, Fig 1C. The concordant set of 327 OTU's that were observed among the four clusters was normalized and scaled, S2 Table.

Pairwise differential abundance analysis of the 327 OTU's was performed among the four tissue groups (NCA, HER2+, ER+, and triple negative breast cancer), S3 Table. Benjamini and Hochberg correction was applied to address multiple testing error and 48 OTU's were observed to be significant (alpha = 0.05) across the pairwise comparisons. A PCA plot of these final 48 OTU's is provided in Fig 1D, demonstrating that the batch factor was accounted for among the normalization and scaling process.

Phyla comparison to Canadian/Irish breast microbiota data

Among 740 TCGA RNA Sequencing tumor samples, nearly half of the identified (48.0%, 157 of 327) reads were derived from *Proteobacteria*. The next most prevalent phyla were *Actinobacteria* (26.3%, 86) and *Firmicutes* (16.2%, 53). The remaining 31 OTUs (9.5%) were from miscellaneous phyla. Moreover, the phylum level composition was consistent among the tumor and NCA tissues among the identified clusters (Fig 2). We present the phylum-level distributions as bean plots, in Fig 2, with the tumors (the distribution densities on the right in purple) and the NCA (the distibution densities on the left in green) clusters. These findings were consistent with Urbaniak, *et al*, shown in the upper right inset in Fig 2.
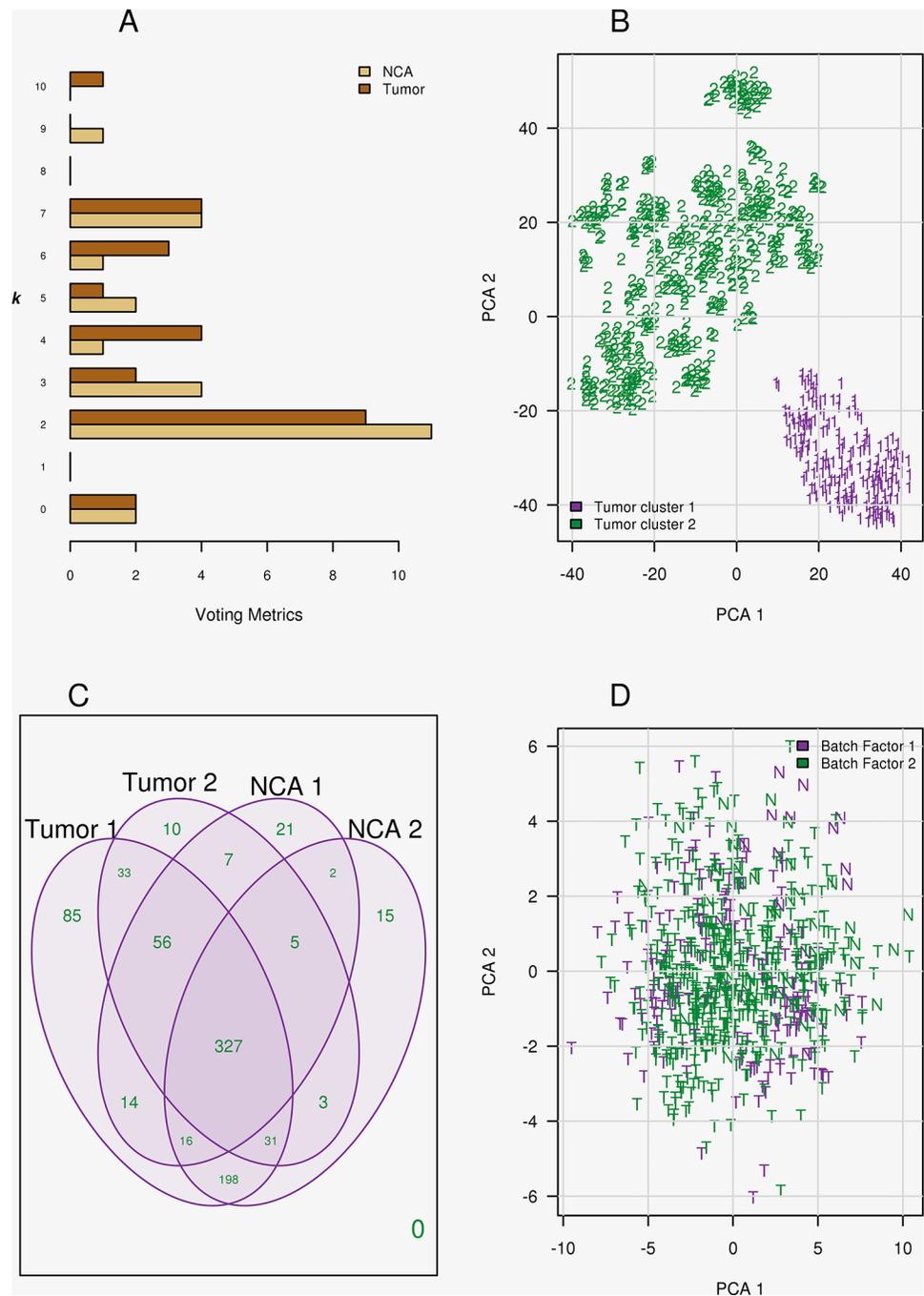
Fig 1. Two distinct microbiota are present in breast tissue samples, both tumor (n = 668) and NCA (n = 72). (A) Optimal k-cluster selection analysis demonstrating a majority decision among the 26 metrics for 2 clusters was concordant for both tissue cohorts identifying a systemic handling bias. (B) The t-SNE projections for the 668 tumors samples, demonstrating tissues demonstrating separation and cohesion among the batches. (C) Filtering was re-applied among the tissue samples and per processing differences. Venn diagram demonstrates a core 327 bacterial OTU's was observed among all samples. (D) Abundance differences were observed for 48 OTU's. A PCA plot of these significant OTU's is presented demonstrating that the batch differences had been accounted for, NCA tissues indicated with an 'N' and tumor samples indicated with a 'T'.
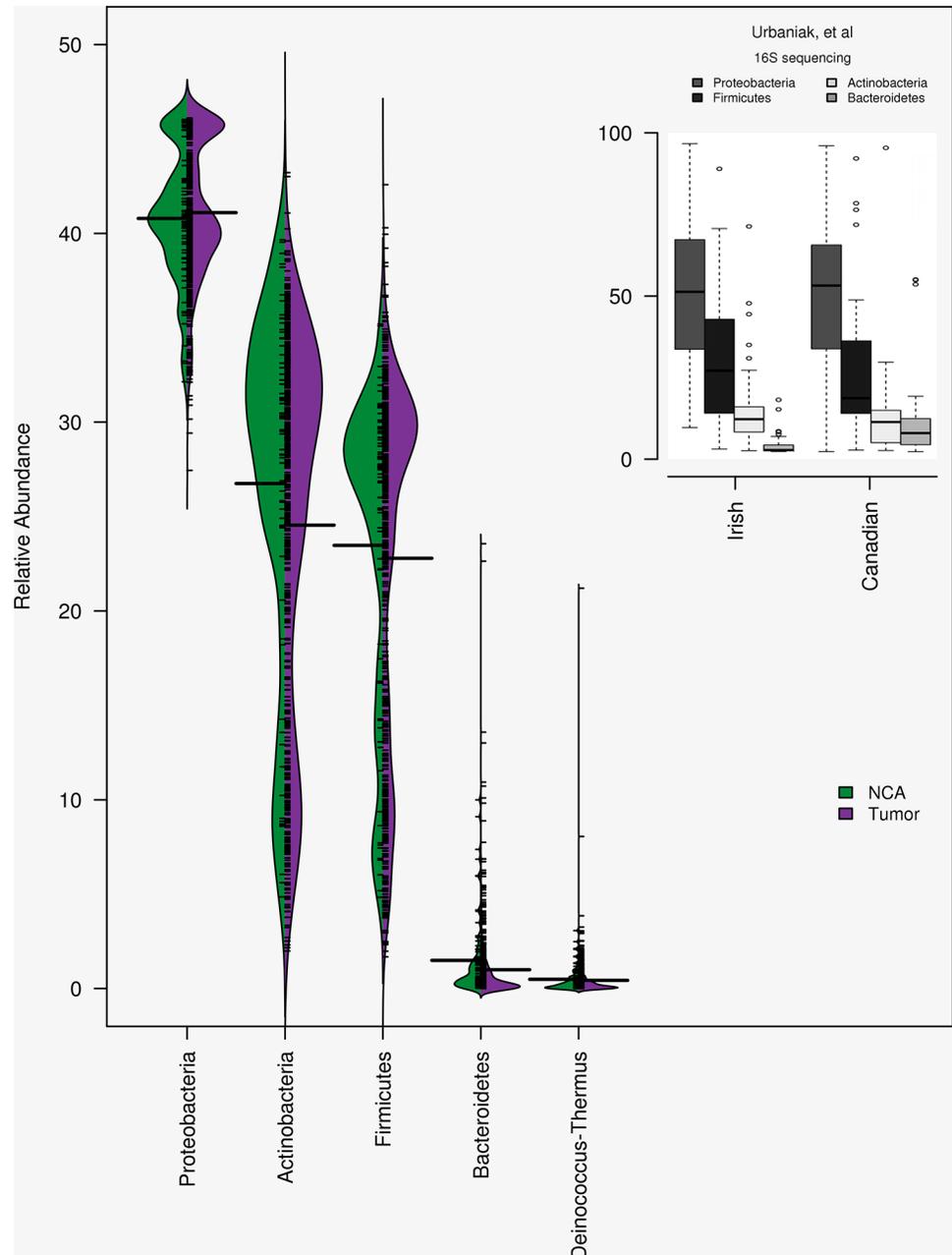
Fig 2. Phylum compositions of the tumor and NCA tissues. The most prevalent phylum in the breast microbiota is *Proteobacteria*, followed by *Actinobacteria* and *Firmicutes*. These observations were consistent among each of the tissue types (tumor shaded in purple NCA shaded in green). These distributions are consistent with the observations by Urbaniak, *et al*, see insert.

We observed 24 of 327 species with an average relative abundance ranging from 0.5–19.3% in the NCA cohort. However, these 24 species account for 85.64% of the overall breast tissue microbiota, across the three breast tumor subtypes as well as the NCA tissue; as presented in Fig 3A. The remainder of the less prevalent microbiota, on average, comprises: 14.7% of the NCA microbiota, 11.1% of the triple negative microbiota, 12.8% of the HER2+ microbiota and 12.5% of the ER+ microbiota.
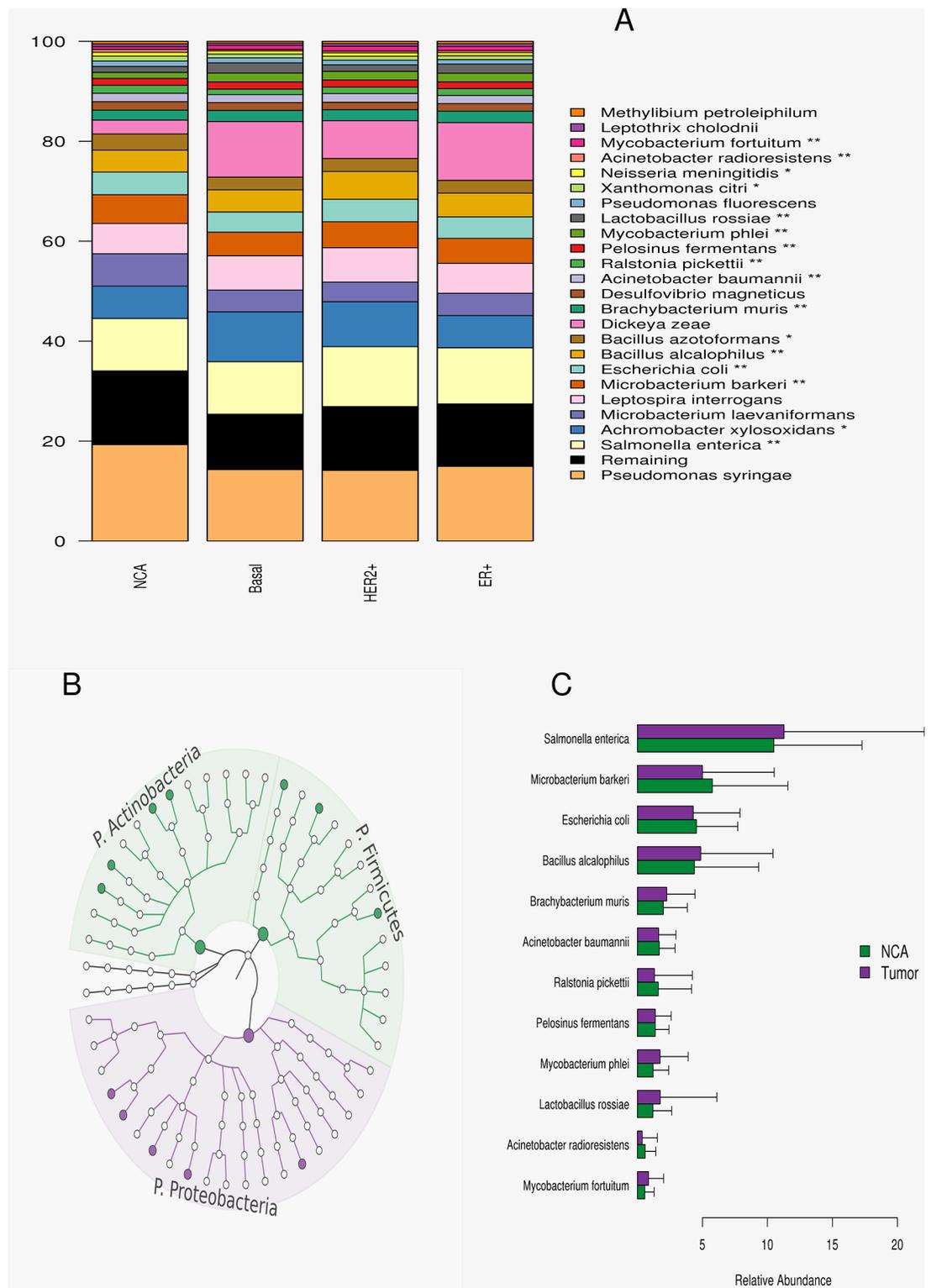
Fig 3. OTU's significantly different among the breast cancer populations. (A) Twenty-four species were observed to have average abundance between 0.5% - 19.3% among the NCA samples. These species are presented as a barplot summary averaged among each subtype. (B) A cladogram of the 48 differentially abundant species, 12 of the 24 prevelent species (average relative abundance greater than 0.5%) were observed to be signicant after Benjamini and Hochberg correction for mutiple testing error. (C) The average relative abundance and standard deviation of the 12

species differing in abundance depicted in the barplots, with the NCA tissue cohort's averages in green and tumor cohort averages in purple.

Significant OTU's. Differential abundance analysis identified 48 OTU's to be significant after correcting for multiple testing errors using Benjamini and Hochberg correction to control the false discovery rate. Nine species were concordantly observed to have absolute abundance differences (fold change) greater than 2 among the breast cancer subtypes, in contrast to the NCA tissues. These 48 species are presented as a cladogram in Fig 3B; the nine species observed to be significantly altered from NCA tissues are presented as shaded circles. The relative abundances of these nine species are also shown in Fig 3C, contrasting the tumor cohort distributions and NCA microbial distributions.

We observed that the prominence of differentially abundant species were comprised from the less prevalent OTU's, as only 12 of the 48 (25%) comprised the 24 more prevalent species. Similarly, the majority (6) of the nine OTU's significantly altered from NCA tissues were also comprised from the less prevalent OTU's. *Mycobacterium fortuitum* and *Mycobacterium phlei*, both known to infect humans were two of the prevalent species observed to be differentially abundant in the tumor samples (Fig 3C). Additionally, we observed an increased presence of *Actinobacteria* (13 of 48 significant OTU's) in the NCA tissue samples; while *Proteobacteria* (20 of the 48 significant OTU's) demonstrated increased presence among the tumor tissues, see S2 Fig.

Microbial association with breast cancer

We investigated whether the microbial composition was associated with alterations in the host expression profiles. EdgeR analysis was performed between the NCA tissue and the breast cancer subtype tissue using host gene expression data. We identified 683 common genes demonstrating logFC greater than two and a Bonferroni corrected p value less than 0.05. We conducted a correlation analysis using the 683 genes and the 48 OTU's. Genes demonstrating a Spearman's correlation (absolute) greater than 0.25 with the OTU's, (S5 Table), were kept for geneset enrichment analysis for the 50 hallmark pathways. Three organisms (*H. influenza*, *N. Subflava*, *and L fleischmannii* were interrogated since they demonstrated sufficient gene correlations for enrichment analysis, with 229, 30, and 58 genes respectively. Fisher's exact test compared each pathway to the reference set of 683 differentially expressed genes. In addition, the gene set enrichment of the pathways for these 683 differentially expressed versus the entire reference set of 16,383 expressed genes, was evaluated as a benchmark for design bias. We observed 14 pathways to be enriched in the 683 differentially expressed genes. Two pathways (G2M checkpoint and E2F targets), were enriched in all four comparisons, while mitotic spindle was enriched in all but *N. Subflava*. However these pathways do not survive any adjustment for multiple testing errors, except in the overall differentially expressed and correlated to *H. influenza*'s composition, where they survive the most stringent Bonferroni correction. Additionally, *L. fleischmannii* remained associated with epithelial mesenchymal transition, after adjusting for the family-wise error rate with the Bonferroni adjustment. The enrichment results are presented as the negative natural log of the p values, in the Fig 4A. We note that the pathway associations with *H. influenza* may follow the general bias of selecting the differentially expressed genes.

We conducted a second association analysis in an effort to reduce the impact of expression bias, by assessing all expressed genes in the 50 hallmark pathways. Mutual information analysis was performed between the microbiota compositional data and expression profiles. The mutual information is empirically estimated based upon the entropies of each variable's
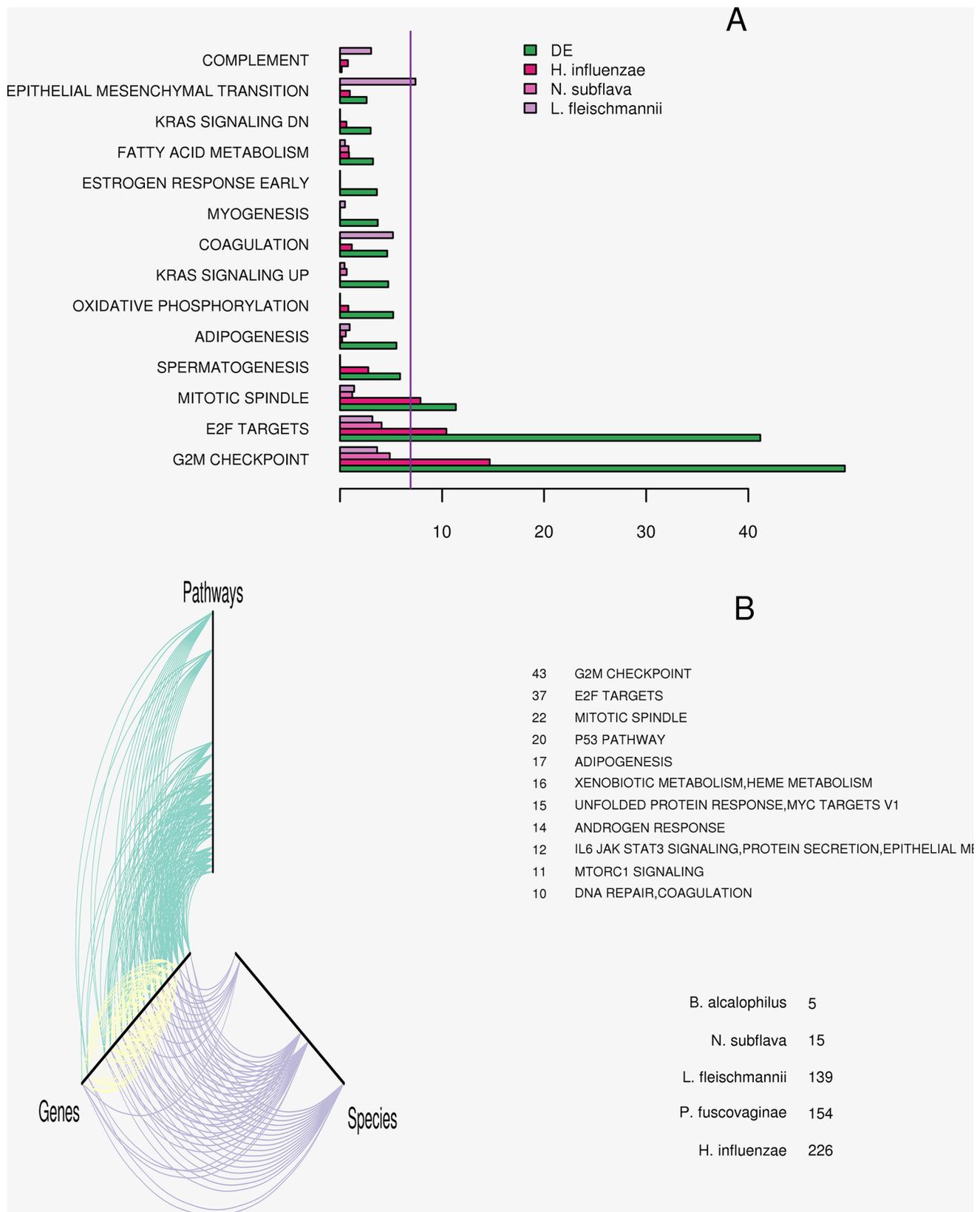
Fig 4. Microbial association with breast cancer. A) Fourteen hallmark pathways which were observed to significantly enriched with differentially expressed genes (in green), versus the universe set of 16,363 genes. Three organisms (*H. influenza*, *N. Subflava*, *and L. fleischmanni*) demonstrated correlation to a sufficient set of genes for geneset enrichment analysis: 229, 30, and 58, genes respectively. *H. influenze* shared enrichment in three of those pathways. Those three pathways (G2M checkpoint, E2F targets, and mitotic spindle), remain enriched, after applying Bonferroni correction (the vertical purple dashed line). *L, Fleischmannii* was also enriched in gene correlations among the epithelial mesenchymal transition pathway, after Bonferroni correction for multiple testing. B) To reduce any bias from differential expression selection the genes in the 50 hallmark pathways were additionally analyzed for mutual information content with the microbial compositional data. A hive plot depicting the network of the 50 pathways and their genes (green), known gene-gene (protein-protein) interactions (yellow), and OTU's-gene associations (purple) is presented. The connectivity of the pathway and OTU's nodes are presented with the text on the right. We again observe the same 3 networks demonstrated the most connectivity and that *H. influenza* as the predominant microbial association.

https://doi.org/10.1371/journal.pone.0188873.g004

distribution and the combined variables distribution. A hive plot is presented in the bottom of Fig 4B, depicting the observed associations among the gene expression, microbiota, and the hallmark pathways. A hive plot collapses a network diagram into 3 axes, where each node represents variables sharing the same connectivity. The more connections (associations) observed the further away from the center origin (x = 0, y = 0, and z = 0). We have provided text detailing the observed association patterns. We observe that same 3 pathways remain the most enriched in genes sharing information with the microbiota, with G2M possessing 43 uniquely associated genes. Similarly *H. influenza* remained the most associated microorganism, with 226 associated genes. While *L. fleischmannii* (139) and epithelial to mesenchymal transitions (12) remained prominently connected features.

16S-rRNA gene sequencing for confirmation

To confirm the presence of microbiota in the TCGA breast samples, for whom the fresh frozen tissue was available, we generated standard 16S-rRNA gene sequencing in a subset of TCGA samples (n = 6) contributed by the Mayo Clinic [18]. In order to control for contamination we included negative controls of the DNA extraction and PCR amplification solutions. Minimal artefactual read evidence was observed for *Geobacillus spp*, as a result of the spiked in DNA enrichment step (14 and 10 reads) and therefore subtracted from our sample cohort (S2 Table and S1 File). The fresh frozen samples were otherwise observed to be predominated by the *Proteobacteria* phylum, consistent with the previously published study of 16S-rRNA gene sequencing with 81 women [10, 55] (S2 Table). Similarly, multiple species were concordantly observed for *Streptococus*, *Lactobacillus*, *and Acinetobacter* [10, 55] (Table 1), while we observed *Gluconacetobacter* as the most prevalent OTU in terms of high read evidence and consistently across samples.

Table 1. Concordantly observed OTU's. The *Proteobacteria* and *Firmicutes* species which were concordantly observed in the 16S-RNA gene and mRNA sequencing data. Additional 16S read data were observed, yet remained uncharacterized at the species level, which could coincide with mRNA observations.

Phyla	Genus	16S and mRNA Confirmed (Species level)	16S Observed (Genus level)
Firmicutes	*Lactobacillus*	*Lactobacillus iners, Lactobacillus helveticus*	*Lactococcus* unclassified
	Streptococcus	*Streptococcus infantis*	*Streptococcus alactolyticus*
	Propionibacterium	*Propionibacterium acnes*	
	Staphylococcus	*Staphylococcus epidermidis*	
Proteobacteria	*Pseudomonas*	*Pseudomonas veronii, Pseudomonas stutzeri*	
	Haemophilus	*Haemophilus parainfluenzae*	
	Acinetobacter	*Acinetobacter johnsonii, Acinetobacter lwoffii*	
	Sphingomonas		*Sphingomonas* unclassified
	Agrobacterium		*Agrobacterium* unclassified

https://doi.org/10.1371/journal.pone.0188873.t001

Discussion

We utilized the unaligned RNA sequence reads available from the TCGA, a multi-institution study and the largest known sequencing cohort available for breast cancer samples. A dual alignment of high quality RNA reads was implemented as false identification was of utmost concern in microbiota identification [52, 53]. Stringent filtering on data sparsity was employed to ensure the identification of the most prevalent OTUs. Additionally, prior to our analysis we removed low quality and/or low RNA yielding samples, as these samples are more susceptible to contaminant amplification [53]. A prominent batch effect was observed accumulating in spurious read evidence for 198 species (Fig 1B and 1C, S1 Table). The ComBat algorithm demonstrated superior performance in adjusting the compositional data normalized with a zero inflated Gaussian model over negative binomial models (data not shown).

Fresh frozen samples, which were from the analogous TCGA patient source, were also assayed with 16S-rRNA gene sequencing (S6 Table). Both datasets confirmed that the most abundant phyla in breast tissues are *Proteobacteria*, *Actinobacteria*, and *Firmicutes* (Fig 2), similar to that reported by smaller 16S-rRNA gene sequencing studies of breast tissue [10, 55]. Comparatively, the mRNA sequencing data appears to be noisier in composition, with over 300 identified species. However, the batch effect which we observed among the TCGA data would likely not have been identified using 16S-rRNA gene sequencing, given the sparsity of the data. While these preparations originate from the same patient, the datasets are derived from distinct laboratory protocols and potentially more importantly different tissue preparations. While the *in silica* processing was different, we were able to concordantly identify ten OTU's at the species level, while an additional three unclassified OTU's were also identified (Table 1, S7 Table). While seemingly unimpressive, the 16S-rRNA gene sequencing approach identified only 30 OTU's, of which only 18 were characterized to named species. The disconnect in tissue preparation and results illustrates the major advantage of implementing the mRNA sequencing approach: the ability to examine two components of the tumor microenvironment using data originating from the same source and *in vitro* processed in the same way.

We identified forty-eight significant OTU's after Benjamini and Hochberg correction for multiple tests. Nine of which had absolute fold changes greater than two, for each breast cancer subtype compared to NCA tissues (S3 Table). Interestingly, the vast majority (75%, 36 of 48) were found to comprise less prevalent OTU's, less than 0.5% on an average in the NCA). While, this may be attributed to the tumor-biased nature of the TCGA data, it could also suggest that breast cancer studies with limited sampling power coupled with sparse microbiome data could be inadequate for accurate signal detection. We observed 24 prevalent (greater than 0.5% on average in the NCA) species across the three breast tumor subtypes and the NCA. These 24 species constituted 85.64% of the overall breast tissue microbiota. Twelve of these twenty-one species were observed to be significantly altered after adjusting for multiple testing errors using Benjamini and Hochberg, across the four 4 breast tissue phenotypes (Fig 3A–3C).

We provide confirmatory evidence that *Escherichia coli* as one of the more prevalent species in the breast tissue and is observed in higher abundance within NCA breast tissues [1, 11, 56]. *Escherichia coli* has been reported to have the ability to inflict DNA damage through its production of colibactin, and could lead to genomic instability which triggers oncogenic process [55]. Similar to published results we observed differential abundance among *Cornebacterium*, *Corynebacterium*, and *Bacillus* and the *enterobaceriaceae*: *E. coli* and *Salmonella enterica [11]*. The *Firmicutes* were represented by 13 significant species, including two *Lactobaccillus spp*. and five *Streptococcus spp*. Another important manner by which breast microbiota could influence oncogenesis is by enhancing the local exposure of breast tissue to estrogen levels. Previously fecal studies demonstrated positive correlations between the abundance of *Streptococcus*

and the presence of β-glucuronidase and/or β-glucosidase enzymes, which cleave the estrogen-glucuronide conjugate and promote recirculation of estrogen [57, 58]. Systemic estrogen levels have been widely recognized to be associated with enhanced breast cancer risk [59] and recent 16S-rRNA gene sequencing experiments reports have associated glucuronidase prevalence in nipple aspirate fluid of breast cancer survivors [20]. The expression profiles for glucosylceramidase beta 2 (GBA2) and the glucuronidase, beta pseudogenes 4 and 9 were positively correlated with *S. pyogenes*. We observed abundance increases specifically for *S. pyogenes* and *L. rossiae* in the tumor samples.

Finally, we interrogated whether any of these OTUs could be associated with host tumor expression profiles. We observed *L. fleischmannii* and to a lesser extent *N. Subflava* to be correlated with tumor gene expression. *Listeria fleischmannii* was more strongly associated with genes involved in the epithelial to mesenchymal transition (Fig 4A). In the same analysis we observed *H. influenza* to be correlated with genes representing pathways fundamental to tumor growth: G2M checkpoint, E2F signaling, and mitotic spindle assembly (Fig 4A and S4 Table). These findings were validated with an unsupervised approach (Fig 4B and S5 Table). *H. influenza* is an opportunistic pathogen which has been demonstrated to elicit an inflammatory immune response and promote tumor growth in murine lung cancer models [60, 61]. We observed an increased presence of *H. influenza* in the NCA samples, suggesting that the organism may more predominantly reside in the surrounding stromal tissue. Further, *H. influenza* has been demonstrated to be susceptible to H_2O_2 production, suggesting that the oxidative tumor environment may be prohibitive to *H. influenza* penetrance and thereby restricting its presence to the surrounding stroma.

Conclusions

In this study, we utilized the largest available collection of TCGA breast cancer RNA sequencing data to examine the microbiota residing in breast tissues. We observed abundance shifts in *Proteobacteria* in tumor samples, while shifts in *Actinobacteria* abundance favored NCA tissues. We observed that the significant shift in microbiota typically involved the less prevalent OTU's, which suggests that under-powered studies are likely to report erroneous results. A semi-supervised approach was implemented to correlate expression profiles with the microbiota compositional data, which revealed *H. influenza* to be significantly correlated with genes in the G2M checkpoint, E2F transcription, and mitotic spindle assembly pathways. Similarly, *L. fleischmannii* was observed to be associated with genes involved with epithelial to mesenchymal transitioning. Additionally, we report correlations between *S. pyogenes* abundance and GUSBP4, GUSBP9, and GPA2 expression levels, which may incriminate the role of *S.pyogenes* in the glucuronidation of estrogen and exposing local breast environment to higher estrogen levels. These findings further support the importance of the host tumor expression in light of the microbial presence within the tumor microenvironment.

Supporting information

S1 File. 16S-rRNA sequencing protocols.
(PDF)

S2 File. Cluster confirmation. Analysis of institutional sampling sources, eliminating Subsequent analysis of tumor data, normalized with DESeq2 and clustered with non-negative matrix factorization.
(PDF)

S3 File. 198 potentially spurious OTU's.
(TXT)

S1 Fig. Two clusters were present in TCGA breast tissue samples, tumor (*n* = 668) and NCA (*n* = 72). (A) Venn diagram of the bacterial OTU's observed among the tumors and NCA, demonstrating similar microbial presence. (B) Optimal *k*-cluster selection analysis demonstrating a majority decision among the 26 metrics for 2 clusters was concordant for both tissue cohorts. (C) A PCA plot of NCA tissues demonstrating separation and cohesion among the smaller second cluster of samples. (D) A perspective plot of the tumor samples demonstrating the sharpness and isolation of tumor samples designated as tumor cluster 1. (E) Silhouette plot for the NCA samples demonstrate a significant proportion, (normal cluster 2, 18 of 72, 0.56) present a distinct batch processing factor which needs to be accounted for. (F) Similarly, the silhouette plot among the tumor samples also demonstrates that a significant proportion (tumor cluster 1,197 of 668, 0.4) also need to be accounted for.
(TIF)

S2 Fig. Parallel sets of the log fold changes observed among the 48 significant OTU's, aggregated among the three major phyla. The observed log fold changes were binned for each of the subtype comparison again NCA tissues.
(TIF)

S1 Table. Retention of the prevalent operational taxonomic units. Table documenting the percent of read count evidence retained as a result of filtering out taxonomic units low abundant OTU's (with less than nine maximum reads across samples) and less prevalent (observed in less than 25% of the sampling population.
(TSV)

S2 Table. SVA batch adjusted microbiota compositional data.
(TSV)

S3 Table. Limma analysis of the adjusted microbiota compositional data.
(TSV)

S4 Table. Spearman correlation of the adjusted microbiota compositional data to RNA Sequencing expression profiles.
(CSV)

S5 Table. Mutual information of the adjusted microbiota compositional data to RNA Sequencing expression profiles.
(TSV)

S6 Table. Microbial read evidence from 16S-rRNA sequencing. Operational taxonomic units for 16S-rRNA sequencing analysis of fresh-frozen tissues samples corresponding with Mayo ER+ contributed TCGA breast cancer samples. The identification of *Geobacillus spp.* is resultant from a DNA spike–in for DNA enrichment in the NEBNext amplification.
(TSV)

S7 Table. Microbial read evidence of species identified concordantly by the two sequencing technologies. Fresh frozen tissue samples from six Mayo Clinic ER+ samples submitted toTCGA were subsequently analyzed with 16S-rRNA gene sequencing for their microbial content. Read evidence for 30 OTU's (31 with the identification of *Geobacillus vulcani* amplification material) were identified, with 18 (60.%) characterized to the species level. Ten of the eighteen (55.6%) characterized species were confirmed with the mRNA sequencing data.
(TSV)

Acknowledgments

Author Contributions

Conceptualization: Kevin J. Thompson, Judy C. Boughey, Krishna R. Kalari.

Data curation: Stephen Johnson, Janet Z. Yao, Sean C. Harrington, Vera J. Suman, Krishna R. Kalari.

Formal analysis: Kevin J. Thompson.

Funding acquisition: James N. Ingle, Matthew P. Goetz, Krishna R. Kalari.

Investigation: Kevin J. Thompson, Krishna R. Kalari.

Methodology: Kevin J. Thompson, Xiaojia Tang, Nicholas Chia, Vera J. Suman, Krishna R. Kalari.

Project administration: Krishna R. Kalari.

Supervision: Krishna R. Kalari.

Visualization: Kevin J. Thompson.

Writing – original draft: Kevin J. Thompson.

Writing – review & editing: Kevin J. Thompson, James N. Ingle, Xiaojia Tang, Nicholas Chia, Patricio R. Jeraldo, Marina R. Walther-Antonio, Karunya K. Kandimalla, Stephen Johnson, Sean C. Harrington, Vera J. Suman, Liewei Wang, Richard L. Weinshilboum, Judy C. Boughey, Jean-Pierre Kocher, Heidi Nelson, Matthew P. Goetz, Krishna R. Kalari.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000; 100(1):57–70. Epub 2000/01/27. PMID: 10647931.

2. Hui L, Chen Y. Tumor microenvironment: Sanctuary of the devil. Cancer Lett. 2015; 368(1):7–13. https://doi.org/10.1016/j.canlet.2015.07.039 PMID: 26276713.

3. Reichman H, Karo-Atar D, Munitz A. Emerging Roles for Eosinophils in the Tumor Microenvironment. Trends Cancer. 2016; 2(11):664–75. https://doi.org/10.1016/j.trecan.2016.10.002 PMID: 28741505.

4. Schwabe RF, Jobin C. The microbiome and cancer. Nat Rev Cancer. 2013; 13(11):800–12. Epub 2013/10/18. https://doi.org/10.1038/nrc3610 PMID: 24132111; PubMed Central PMCID: PMC3986062.

5. Chen Y, Wei J. Identification of Pathogen Signatures in Prostate Cancer Using RNA-seq. PLoS One. 2015; 10(6):e0128955. https://doi.org/10.1371/journal.pone.0128955 PMID: 26053031; PubMed Central PMCID: PMCPMC4460021.

6. Pevsner-Fischer M, Tuganbaev T, Meijer M, Zhang SH, Zeng ZR, Chen MH, et al. Role of the microbiome in non-gastrointestinal cancers. World J Clin Oncol. 2016; 7(2):200–13. https://doi.org/10.5306/wjco.v7.i2.200 PMID: 27081642; PubMed Central PMCID: PMCPMC4826965.

7. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. 2012; 22(2):299–306. https://doi.org/10.1101/gr.126516.111 PMID: 22009989; PubMed Central PMCID: PMCPMC3266037.

8. Zackular JP, Rogers MA, Ruffin MTt, Schloss PD. The human gut microbiome as a screening tool for colorectal cancer. Cancer Prev Res (Phila). 2014; 7(11):1112–21. https://doi.org/10.1158/1940-6207.CAPR-14-0129 PMID: 25104642; PubMed Central PMCID: PMCPMC4221363.

9. Erb-Downward JR, Thompson DL, Han MK, Freeman CM, McCloskey L, Schmidt LA, et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. PLoS One. 2011; 6(2):e16384. https://doi.org/10.1371/journal.pone.0016384 PMID: 21364979; PubMed Central PMCID: PMCPMC3043049.

10. Hieken TJ, Chen J, Hoskin TL, Walther-Antonio M, Johnson S, Ramaker S, et al. The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. Sci Rep. 2016; 6:30751. https://doi.org/10.1038/srep30751 PMID: 27485780; PubMed Central PMCID: PMCPMC4971513.

11. Urbaniak C, Gloor GB, Brackstone M, Scott L, Tangney M, Reid G. The Microbiota of Breast Tissue and Its Association with Breast Cancer. Appl Environ Microbiol. 2016; 82(16):5039–48. https://doi.org/10.1128/AEM.01235-16 PMID: 27342554; PubMed Central PMCID: PMCPMC4968547.

12. Xuan C, Shamonki JM, Chung A, Dinome ML, Chung M, Sieling PA, et al. Microbial dysbiosis is associated with human breast cancer. PLoS One. 2014; 9(1):e83744. Epub 2014/01/15. https://doi.org/10.1371/journal.pone.0083744 PMID: 24421902; PubMed Central PMCID: PMC3885448.

13. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin. 2015; 65(1):5–29. Epub 2015/01/07. https://doi.org/10.3322/caac.21254 PMID: 25559415.

14. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016; 66(1):7–30. https://doi.org/10.3322/caac.21332 PMID: 26742998.

15. Kenney NJ, Dickson RB. Growth factor and sex steroid interactions in breast cancer. J Mammary Gland Biol Neoplasia. 1996; 1(2):189–98. PMID: 10887492.

16. Anderson WF, Rosenberg PS, Prat A, Perou CM, Sherman ME. How many etiological subtypes of breast cancer: two, three, four, or more? J Natl Cancer Inst. 2014; 106(8). https://doi.org/10.1093/jnci/dju165 PMID: 25118203; PubMed Central PMCID: PMCPMC4148600.

17. Gierach GL, Burke A, Anderson WF. Epidemiology of triple negative breast cancers. Breast Dis. 2010; 32(1–2):5–24. https://doi.org/10.3233/BD-2010-0319 PMID: 21965309; PubMed Central PMCID: PMCPMC4684941.

18. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. https://doi.org/10.1038/nature11412 PMID: 23000897; PubMed Central PMCID: PMCPMC3465532.

19. Fuhrman BJ, Feigelson HS, Flores R, Gail MH, Xu X, Ravel J, et al. Associations of the fecal microbiome with urinary estrogens and estrogen metabolites in postmenopausal women. J Clin Endocrinol Metab. 2014; 99(12):4632–40. https://doi.org/10.1210/jc.2014-2222 PMID: 25211668; PubMed Central PMCID: PMCPMC4255131.

20. Chan AA, Bashir M, Rivas MN, Duvall K, Sieling PA, Pieber TR, et al. Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors. Sci Rep. 2016; 6:28061. https://doi.org/10.1038/srep28061 PMID: 27324944; PubMed Central PMCID: PMCPMC4914981.

21. Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013; 31(9):814–21. https://doi.org/10.1038/nbt.2676 PMID: 23975157; PubMed Central PMCID: PMCPMC3819121.

22. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database (Oxford). 2011; 2011:bar026. Epub 2011/09/21. https://doi.org/10.1093/database/bar026 PMID: 21930502; PubMed Central PMCID: PMC3263593.

23. Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford). 2014;2014. Epub 2014/10/01. https://doi.org/10.1093/database/bau093 PMID: 25267794; PubMed Central PMCID: PMC4178372.

24. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31(2):166–9. Epub 2014/09/28. https://doi.org/10.1093/bioinformatics/btu638 PMID: 25260700; PubMed Central PMCID: PMC4287950.

25. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. Epub 2009/03/18. https://doi.org/10.1093/bioinformatics/btp120 PMID: 19289445; PubMed Central PMCID: PMC2672628.

26. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13(2):204–16. Epub 2012/01/31. https://doi.org/10.1093/biostatistics/kxr054 PMID: 22285995; PubMed Central PMCID: PMC3297825.

27. Mahoney DW, Therneau TM, Anderson SK, Jen J, Kocher JP, Reinholz MM, et al. Quality assessment metrics for whole genome gene expression profiling of paraffin embedded samples. BMC Res Notes. 2013; 6:33. Epub 2013/01/31. https://doi.org/10.1186/1756-0500-6-33 PMID: 23360712; PubMed Central PMCID: PMC3626608.

28. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014; 15(3):R46. Epub 2014/03/04. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807; PubMed Central PMCID: PMC4053813.

29. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016. https://doi.org/10.1038/nbt.3601 PMID: 27454739.

30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.

31. Jeraldo P, Kalari K, Chen X, Bhavsar J, Mangalam A, White B, et al. IM-TORNADO: a tool for comparison of 16S reads from paired-end libraries. PLoS One. 2014; 9(12):e114804. https://doi.org/10.1371/journal.pone.0114804 PMID: 25506826; PubMed Central PMCID: PMCPMC4266640.

32. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009; 75(23):7537–41. https://doi.org/10.1128/AEM.01541-09 PMID: 19801464; PubMed Central PMCID: PMCPMC2786419.

33. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006; 72(7):5069–72. https://doi.org/10.1128/AEM.03006-05 PMID: 16820507; PubMed Central PMCID: PMCPMC1489311.

34. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013; 10(12):1200–2. Epub 2013/10/01. https://doi.org/10.1038/nmeth.2658 PMID: 24076764; PubMed Central PMCID: PMC4010126.

35. Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. PLoS Comput Biol. 2012; 8(12): e1002808. https://doi.org/10.1371/journal.pcbi.1002808 PMID: 23300406; PubMed Central PMCID: PMCPMC3531975.

36. Oksanen J, Blanchet F, Guillaume Kindt R, Legendre P, Minchin P, R., O'Hara R, B., et al. vegan: Community Ecology Package. R Foundation for Statistical Computing: Vienna; 2015.

37. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012; 9(8):811–4. Epub 2012/06/13. https://doi.org/10.1038/nmeth.2066 PMID: 22688413; PubMed Central PMCID: PMC3443552.

38. Donaldson J. tsne: T-distributed Stochastic Neighbor Embedding for R (t-SNE). R Foundation for Statistical Computing: Vienna; 2012.

39. Charrad M, Ghazzali N., Boiteau V., and Niknafs A. NbClust: An {R} Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software. 2014; 61:36.

40. Maechler M, Rousseeuw P, Struyf A, Hubert M. cluster: Cluster Analysis Extended Rousseeuw et al. R Foundation for Statistical Computing: Vienna; 2015.

41. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20:53–65.

42. Leek JT, Johnson W. Evan, Parker Hilary S., Fertig Elana J., Jaffe Andrew E., Storey John D. sva: Surrogate Variable Analysis. 3.14.0 ed. Vienna, Austria: R Foundation for Statistical Computing; 2015.

43. Smyth G, Ritchie M, Silver J, Wettenhall J, Thorne N, McCarthy D, et al. limma: Linear Models for Microarray Data. R Foundation for Statistical Computing: Vienna; 2015.

44. Chen Y, Lun A, McCarthy D, Zhou X, Robinson M, Smyth G. edgeR: Empirical analysis of digital gene expression data in R. R Foundation for Statistical Computing: Vienna; 2015.

45. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1–13. https://doi.org/10.1093/nar/gkn923 PMID: 19033363; PubMed Central PMCID: PMCPMC2615629.

46. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1(6):417–25. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021; PubMed Central PMCID: PMCPMC4707969.

47. Jin Y, Peng H., Wang L., Fronza R., Liu Y. Nardini C. BUS: Gene network reconstruction. 1.24.0 ed: R Foundation for Statistical Computing; 2010.

48. Kampstra P. beanplot: Visualization via Beanplots (like Boxplot/Stripchart/ViolinPlot). R Foundation for Statistical Computing: Vienna; 2014.

49. Hanson BA. HiveR: 2D and 3D Hive Plots for R. 0.2.55 ed: R Foundation for Statistical Computing; 2016.

50. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. Compact graphical representation of phylogenetic data and metadata with GraPhlAn. PeerJ. 2015; 3:e1029. Epub 2015/07/15. https://doi.org/10.7717/peerj.1029 PMID: 26157614; PubMed Central PMCID: PMC4476132.

51. Schulz A. epade: Easy Plots. 0.3.8 ed. Vienna, Austria: R Foundation for Statistical Computing; 2013.

52. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014; 2:e675. Epub 2014/11/27. https://doi.org/10.7717/peerj.675 PMID: 25426337; PubMed Central PMCID: PMC4243333.

53. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014; 12:87. Epub 2014/11/13. https://doi.org/10.1186/s12915-014-0087-z PMID: 25387460; PubMed Central PMCID: PMC4228153.

54. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017; 130(4):453–9. https://doi.org/10.1182/blood-2017-03-735654 PMID: 28600341.

55. Urbaniak C, Cummins J, Brackstone M, Macklaim JM, Gloor GB, Baban CK, et al. Microbiota of human breast tissue. Appl Environ Microbiol. 2014; 80(10):3007–14. Epub 2014/03/13. https://doi.org/10.1128/AEM.00242-14 PMID: 24610844; PubMed Central PMCID: PMC4018903.

56. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5):646–74. Epub 2011/03/08. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230.

57. Flores R, Shi J, Fuhrman B, Xu X, Veenstra TD, Gail MH, et al. Fecal microbial determinants of fecal and systemic estrogens and estrogen metabolites: a cross-sectional study. J Transl Med. 2012; 10:253. https://doi.org/10.1186/1479-5876-10-253 PMID: 23259758; PubMed Central PMCID: PMCPMC3552825.

58. Flores R, Shi J, Gail MH, Gajer P, Ravel J, Goedert JJ. Association of fecal microbial diversity and taxonomy with selected enzymatic functions. PLoS One. 2012; 7(6):e39745. https://doi.org/10.1371/journal.pone.0039745 PMID: 22761886; PubMed Central PMCID: PMCPMC3386201.

59. Baumgarten SC, Frasor J. Minireview: Inflammation: an instigator of more aggressive estrogen receptor (ER) positive breast cancers. Mol Endocrinol. 2012; 26(3):360–71. https://doi.org/10.1210/me.2011-1302 PMID: 22301780; PubMed Central PMCID: PMCPMC3286192.

60. Chang SH, Mirabolfathinejad SG, Katta H, Cumpian AM, Gong L, Caetano MS, et al. T helper 17 cells play a critical pathogenic role in lung cancer. Proc Natl Acad Sci U S A. 2014; 111(15):5664–9. https://doi.org/10.1073/pnas.1319051111 PubMed PMID: WOS:000334288600056. PMID: 24706787

61. Moghaddam SJ, Li HH, Cho SN, Dishop MK, Wistuba II, Ji L, et al. Promotion of Lung Carcinogenesis by Chronic Obstructive Pulmonary Disease-Like Airway Inflammation in a K-ras-Induced Mouse Model. Am J Resp Cell Mol. 2009; 40(4):443–53. https://doi.org/10.1165/rcmb.2008-0198OC PubMed PMID: WOS:000264663300007. PMID: 18927348