Community Dynamics of Fungi During the Progression of Human Non-Small-Cell Lung Cancer

Yiming Zhao  
Central South University

Junqi Yi  
Central South University

Juanjuan Xiang  
Central South University

Wei Jia  
Hong Kong Baptist College: Hong Kong Baptist University

Anqi Chen  
Central South University

Liyu Chen  
Central South University

Leliang Zheng  
Central South University

Wen Zhou  
Central South University

Minghua Wu  
Central South University

Zheng Yu  
Central South University  
Zheng Yu (✉️ yuzheng@csu.edu.cn)  
https://orcid.org/0000-0001-7366-3946

Jingqun Tang  
Central South University

Research

Keywords: non-small-cell lung cancer, Fungi, Metagenomic sequence, Alternaria arborescens

DOI: https://doi.org/10.21203/rs.3.rs-616977/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Background: Lung cancer is the most common cancer worldwide. The correlation between lung microbiota dysbiosis and lung cancer has been explored. However, the characteristics of the lung mycobiome in lung cancer patients has not been fully examined. In this study, we performed metagenomic sequencing on bronchoalveolar lavage (BAL) from 24 patients with non-small-cell lung cancer (NSCLC) and 14 non-cancerous controls with benign mass.

Results: We observed greater fungal diversity and connectivity in NSCLC patients compared to non-cancer controls. Network analysis revealed that the sum of degrees in NSCLC was more numerous, suggesting the connections of fungal communities were more active. The community networks were reconstructed in 4 main ecological modules. The module and network hubs included *Vanderwaltozyma polyspora*, *Apiotrichum porosum*, *Lobosporangium transversal*, *Capronia epimyces* and *Pseudomicrostroma glucosiphilum*. Random forest analysis and DESeq2 were performed to screen out distinct species in two separate groups and 3 fungi *Alternaria arborescens*, *Eremoyces bilateralis* and *Malassezia restricta* were identified as potential diagnosis biomarkers for NSCLC patients. In these 3 fungi, the relative abundance of *Alternaria arborescens* was gradually increasing as NSCLC progressed.

Conclusions: Taken together, our study innovatively proposes that changes in the structure of fungal communities in lung cancer may be closely related to lung carcinogenesis as well as its progression and the distinct species listed above could be used to detect and monitor NSCLC.

Introduction

Lung cancer is the most common cancer worldwide. Although smoking is the greatest risk factor for developing lung cancer, more than 25% of lung cancer patients have never smoked [1]. Microbial exposure exerts a long-lasting effect on the carcinogenesis of lung cancer. Healthy lungs that were previously thought to be sterile are now known to harbor a dynamic ecosystem including bacteria, fungi and viruses [2]. Recent studies have revealed that the lung has a distinct microbiome and that the lower respiratory tract microbiota were implicated in lung carcinogenesis [3].

Although there was less microbial biomass when compared to the intestinal tract, the microbiome of the lung detected by next-generation sequencing technologies displayed considerable diversity. The predominant phyla in the healthy lung were *Bacteroidetes* and *Firmicutes* [4]. A significant enrichment of *Granulicatella*, *Abiotrophia*, and *Streptococcus* at the genus level has been observed in lung cancer patients [3]. Our previous study showed that although the abundance of some species was higher in lung cancer patients compared to non-cancer controls, the gross composition of microbiota did not change dramatically [5]. Similar to the bacterial microbiota, mycobiota, which include fungi, also affect lung physiology and pathology. Unlike the human microbiome, the human mycobiome has been poorly characterized. Fungi are eukaryotic microorganisms found in the environment that enter the lungs when a person inhales those which are airborne. The incidence of fungal infection has increased in
immunosuppressed populations [6]. Individual members of the mycobiome may play a beneficial or pathogenic role [7]. Fungal communities have been studied by PCR amplification and high-throughput sequencing of their small-subunit ribosomal RNA gene [8, 9]. Fungal genera that have been detected in the pulmonary mycobiota include *Candida*, *Malassezia*, *Neosartorya*, *Saccharomyces*, and *Aspergillus* [7]. *Cladosporium* spp., *Aspergillus* spp. and *Penicillium* spp. dominate fungal genera in oral and nasal cavities [10]. The presence of *Fusarium* and *Cryptococcus* in lungs but not in the oral cavity suggested that the lower airways harbor fungal communities distinct from the upper airways [11]. However, the roles of fungi in lung cancer have not been explored.

In this study, we collected bronchoalveolar lavage (BAL) fluid samples using bronchoscopy. Considering the low density and high host DNA contamination in BAL samples, we performed shotgun metagenomic sequencing to characterize the fungal communities in lungs and then focused on the fungal composition in non-small cell lung cancer (NSCLC) patients.

**Methods**

**Study population and sample selection**

For metagenomic sequencing, BAL were obtained from 38 individuals, including 24 newly diagnosed NSCLC patients and 14 non-cancerous controls with benign mass (3 thymoma patients, 11 patients with benign lung mass). All cases enrolled in this study were collected at the second Xiangya hospital, Central South University, China. These patients underwent bronchoscopy for diagnosis or therapy. The 3 patients with thymoma underwent postoperative aspiration to keep the airway unobstructed. There were no adverse clinical events related to sampling. The clinical characteristics of the patients are presented in Table 1. We collected the BAL from the bronchoscopic aspirate. The 38 samples were obtained under sterile conditions by instillation and aspiration of 20 ml of 0.9% NaCl from the bronchoscope. These patients had not been treated with antibiotics. Patients with clinical evidence of infection, sepsis or active tuberculosis were excluded. The patients were informed of the sample collection and signed informed consent forms. The collection and use of samples were approved by the ethical review committees of the second Xiangya Hospital, Central South University.
| Clinical characteristics of patients |
|-------------------------------------|
| Patients (n = 38)                    |
| NSCLC patients                      | 24 (63.2%) |
| Control Benign Mass Patients        | 14 (36.8%) |
| Genders (n = 38)                    |
| Male                                | 20 (52.6%) |
| Female                              | 18 (47.4%) |
| Age (n = 38)                        |
| ≤ 50                                | 16 (42.1%) |
| >50                                 | 22 (57.9%) |
| Smoking status (n = 38)             |
| Smoker                              | 6 (15.8%)  |
| Non-smoker                          | 32 (84.2%) |
| TNM staging (n = 24)                |
| T1aN0M0                             | 2 (8.3%)   |
| T1bN0M0                             | 2 (8.3%)   |
| T1cN0M0                             | 7 (29.2%)  |
| T2aN0M0                             | 9 (37.5%)  |
| T2bN1M0                             | 1 (4.2%)   |
| T3N1M0                              | 1 (4.2%)   |
| T4N0M0                              | 1 (4.2%)   |
| Tis                                 | 1 (4.2%)   |
Table 2  
Cancer stage

| Staging and Grading Cancer | Quantitative digital |
|----------------------------|----------------------|
| non-cancer controls       | 0                    |
| T1a                       | 1                    |
| T1a                       | 2                    |
| T1b                       | 3                    |
| T1c                       | 4                    |
| T2a                       | 5                    |
| T2b                       | 6                    |
| T2c                       | 7                    |
| T3a                       | 8                    |
| T3b, T2bN1                | 9                    |
| T3c, T3N1                 | 10                   

Metagenomic sequencing and taxonomic classification

The metagenomic sequencing was performed using the paired-end sequencing method on the Illumina platform (San Diego, CA, USA). Briefly, DNA was sheared by ultrasonication (Covaries, Woburn, MA). The sheared DNA fragments were end-repaired (DNA End Repair Mix) at 20°C for 30 min. The DNA fragments were purified by QIAquick PCR Purification Kit (Qiagen) and A-tailed using A-Tailing Mix. Libraries were checked using Bioanalyzer 2100 (Agilent) and quantified using the ABI StepOnePlus Real-Time PCR System. Libraries were sequenced on an Illumina platform. Raw sequences were processed to remove low-quality sequences using fastp [12] (Version 0.21.0), and fastuniq [13] (Version 1.1.0) to remove duplicates in paired short DNA sequence reads in a FASTQ format. Human sequences were filtered out using the human reference genome (hg37) using Bowtie2 [14] (Version 2.3.5).

The remaining high-quality reads were used for taxonomic classification by Kraken2 [15] (Version 2.0.7), utilizing the custom Kraken fungal database with default settings. The read counts table of several levels (e.g., phylum, class, order, family, genus, species) were rarefied to the minimum read counts to reduce the effects of uneven sampling in the cohort by using R (Version 3.6.1) and package picante [16] (Version 1.8.2).

Fungal diversity in the cohort

To evaluate the sequencing depth and species richness in all subjects in the cohort, rarefaction analysis was performed by using R (version 3.6.1) and package picante (Version 1.8.2). A diversity index takes into consideration the number of species, the number of individuals of different species in a sample, or
habitat or a community [17]. To estimate the fungal diversity, the Shannon–Wiener index, Gini-Simpson index and Pielou index were also calculated. In addition, the Bray-Curtis dissimilarity indices between samples at the species level were calculated by using R package vegan [18]. And based on it, the permutational multivariate analysis of variance (PERMANOVA) and principal coordinate analysis (PCoA) were performed to estimate the between sample (β) diversity.

Network construction and analysis

To explore the different fungal correlation between NSCLC and non-cancer controls, the networks were constructed based on relative abundances of different species. In order to avoid unreliable results, the FastSpar [19] (Version 1.0) was also used to realize the Sparcc network efficiently. Only robust (|r| > 0.7) and statistically significant (P-value < 0.01) correlations were considered in network analyses. The covariation was measured across all relative abundance of filtered fungi to create each network. Degree is the measure of the total number of edges connected to a particular vertex. The networks were constructed by R package igraph (Version 1.2.6). And networks were graphed by using Cytoscape (Version 3.8) and gephi (Version 0.9.2).

Detection of modules and identification of node roles

We characterized network modularity for networks created in this study. A module is a group of nodes (i.e., species) that are highly connected within the group with few connections outside the group [20]. Based on connectivity and connectors connecting different modules were identified in co-occurrence network using the greedy modularity optimization method by the package ‘igraph’ (Version 1.2.6) in R environment [21]. The connectivity of each node was determined based on its within-module connectivity (Zi) and among-module connectivity (Pi) [22] which were then used to classify the nodes based on the topological roles they played in the network. Node topologies are organized into four categories: module hubs (highly connected nodes within modules, Zi > 2.5), network hubs (highly connected nodes within entire network, Zi > 2.5 and Pi > 0.62), connectors (nodes that connect modules, Pi > 0.62) and peripherals (nodes connected in modules with few outside connections, Zi < 2.5 and Pi < 0.62). In the field of ecology, a species can be classified as a generalist or a specialist which is a way to identify what kinds of food and habitat resources it relies on to survive. Generalists can eat a variety of foods and thrive in a range of habitats. Specialists, on the other hand, have a limited diet and stricter habitat requirements. In the network analysis, peripherals might represent specialists which only associate with few species in networks, whereas module hubs and connectors were close to generalists which could associate with lots of species and network hubs represent super-generalists [21, 23, 24].

Bioinformatic analysis

On the R platform, metagenomic features (i.e., taxonomic) were analyzed statistically. The top 15 species and the top thirty genera were found to account for the majority of fungi. The different species were identified by R package DESeq2 [25] (Version 1.26.0). In order to acquire the best discriminant performance of taxa between NSCLC and non-cancer controls, we used the relative abundances of fungal taxa at the species level to achieve machine learning using the Python library ‘sklearn’. The Random
forests approach is one of the most robust ensemble machine learning methods for classification based on the construction of many single classification trees. Lists of taxa ranked by Random forests were determined over 100 iterations. The accuracy of model was identified using 10-fold cross-validation implemented with the ‘cross_val_score’ function in the Python library ‘sklearn’. After adjusting parameters, we retained the top 20 most important species to make the accuracy of the Random forests model improve to 76.7%.

Results

Greater fungal diversity in patients with NSCLC

In this study, only the fungal community was taken into consideration. The fungal composition in each sample was shown in Fig. 1a. The majority of fungal read counts in the cohort were dominated by the species Lasiodiplodia theobromae and Malassezia globosa, representing 64.62% and 11.83% of the fungi, respectively, followed by Grosmannia clavigera (4.7%), Botrytis fragariae (2.8%) and Aspergillus flavus (2.3%). Thus, about 85% of fungal read counts were covered by the top five most abundant species in both non-cancer controls and NSCLC patients (Figure S1a, Additional file1). Among the five species, Botrytis fragariae was more abundant in NSCLC patients than non-cancer controls (Figure S1b-c, Additional file1). The species rarefaction curve for each sample was performed and was found to approach saturation, indicating that the sequencing depth was adequate (Figure S1d, Additional file1). In addition, we found that NSCLC samples represented by red rarefaction curves exhibited higher species richness. Indeed, the higher fungal α diversity (within-habitat diversity) in NSCLC was confirmed by the greater Shannon-Wiener index and Gini-Simpson index at the species level significantly (p < 0.05) (Fig. 1b).

The different fungal β diversity (between-habitat diversity) between non-cancer controls and NSCLC can also be visualized using statistical significance with values (p = 0.0012) (Fig. 1c). Furthermore, we estimated the differences in β-diversity among different groups based on the Bray–Curtis distance which indicated that the NSCLC communities significantly showed a higher β-diversity and higher dispersion (p = 0.003) (Figure S1e, Additional file1). Thus, greater fungal composition diversity corresponded positively with lung cancer in this cohort.

Greater fungal connectivity in patients with NSCLC

To study the community differences of fungi between the two groups, we constructed two networks of non-cancer controls and NSCLC (Fig. 2a). The degree distributions of the fungal co-occurrence networks conformed to the power-law distribution, indicating that the fungal community was constructed in a non-random way (Figure S2a-b, Additional file2). The majority of covariation was positive in both networks. The classes that made up the two networks were somewhat different. The top 5 species of most fungal degrees, which suggested the top 5 highest co-occurrence rates, in the NSCLC group belonged to the classes, Dothideomycetes (includes top 1 and top 4 species), Sordariomycetes, Leotiomycetes and Lecanoromycetes. However, in the non-cancer controls, the top species of most fungal degrees belonged
to the classes *Sordariomycetes*, *Leotiomycetes* (include top 2 to top 4 species), and *Lecanoromycetes* (Figure S2c-d, Additional file2). In the network analysis, the fungal connectivity of NSCLC was significantly higher than non-cancer controls (*p* < 0.0001) (Fig. 2b-c). Regarding the NSCLC network, the majority of co-occurrence relationships was in module1 (63.35%) and module2 (27.21%) (Fig. 2d).

**Fungal alpha diversity increases as cancer progresses**

As shown in the results above, the diversity of the NSCLC was significantly different from that of the non-cancer controls, and there were more fungal associations in the NSCLC. In order to further explore the diversity changes in cancer progression, we did regression analysis of indices which represent α diversity and cancer stage (from 0 to 10). Interestingly, all results exhibited the significant tendency that α diversity increased with cancer stage (all *p* < 0.01) (Fig. 2e-g).

**Distinct fungal abundance in different NSCLC module**

In order to further explore the relationship between the co-occurring networks of fungal communities in the NSCLC patients, we focused on the 4 main modules which had the most degree (Fig. 3a). In Fig. 3b, the composition of different circles represents the relative abundance of the class-level species in the different modules. The fungal composition within each module was obviously different. The 'Dominant fungi' with the highest relative abundance in the four modules were *Dothideomycetes*, *Leotiomycetes*, *Eurotiomycetes* and *Saccharomycetes*, respectively. *Malasseziomycetes* did not appear in module 2 but was detected in module1.

In module 2, which had the most abundant connectivity (degree) for NSCLC, *Leotiomycetes* had the highest relative abundance, followed by *Eurotiomycetes* and *Dothideomycetes*. But in module 1, the relative abundance of *Dothideomycetes* was highest than half of all. So, module 1 may be the one most affected by environmental factors (changes in the lung airway environment). Next, we performed regression analysis of the relative abundance for the four modules against cancer stage (Fig. 3c). It can be seen that except for the module 2, the relative abundance of the remaining three modules increased, and module 1 was the best fit model in linear regression analysis (*R²* = 0.151, *p* < 0.002).

**Hubs of module and network as putative keystone taxa**

To assess possible topological roles of taxa in the networks, we classified nodes into 4 categories: peripherals, connectors, module hubs and network hubs based on their within-module connectivity (*Zi*) and among-module connectivity (*Pi*) (Fig. 3d). The module hubs and connectors were close to generalists and network hubs as super-generalists, suggesting that they have more communication with the whole community. Among them, the network hub was *Pleurotus ostreatus*, and that the module hubs included five species which were *Vanderwaltozyma polyspora*, *Apiotrichum porosum*, *Lobosporangium transversal*, *Capronia epimyces* and *Pseudomicrostroma glucosiphilum*. We also found that half of these hubs were significantly higher in the NSCLC group than the non-cancer controls (*p* < 0.05) (Figure S3, Additional file3). In addition, the NSCLC network contained 12 connectors.
Different fungal abundance may lead to changes in fungal community

In order to explore the altered performance of the whole fungal community of NSCLC at the species level, we used two different methods to infer the distinct species between the NSCLC and the non-cancer controls. 27 distinct species were screened out by Deseq2 (Fig. 4a) and a heatmap graph colored with blue was made to show the difference of their relative abundance (Fig. 4b). Subsequently, to reduce the randomness and inaccuracy of the algorithm, we also used a Random Forest algorithm to analyze the same data and screened out 20 species (Figure S4a, Additional file4) and colored with orange in Fig. 4b. Figure 4c showed that there were 3 distinct fungi (Alternaria arborescens, Eremoyces bilateralis and Malassezia restricta) screened out by both Random Forest and DESeq2, which may have the potential to be diagnosis biomarkers for NSCLC patients and the relative abundances of them were shown in Fig. 4b with green. Next, we took the intersection of the results of Random Forest, DESeq2 and did an analysis of regression (Figure S4b, Additional file4). Alternaria arborescens was detected that its relative abundance was gradually increasing as NSCLC progressed (Fig. 4c).

Discussion

In our study, we first found that the composition of the lung airway fungal community was significantly different between NSCLC and non-cancer controls and that the fungal diversity was significantly increased in the NSCLC group (Fig. 1). In the field of microbes, correlation network analysis provides a powerful research tool for the development of microbial ecology and helps to find and identify key microorganisms [26]. Network analysis can reveal non-random species co-occurrence patterns in the microbial flora [27], so that the direct interaction or niche sharing characteristics between species can be compared. When we focused on the general level of NSCLC and non-cancer control fungal ecology, we saw that compared with the non-cancer controls, the fungal communities in NSCLC were reorganized, their connections became more numerous.

Fungi are common in indoor environments and they use spores to spread and colonize. In various ecosystems, fungi exist as pathogens, mutualists and decomposers. From the BAL collected from the 38 patients, the fungal species with the highest relative abundance in lungs were Lasiodiplodia theobromae and Malassezia globosa. Lasiodiplodia theobromae and Malassezia are naturally found in plants and skin and we hypothesized that these fungi were inhaled. The prevalence of these fungi may represent regional bias. In the NSCLC group, we divided the fungal flora into 12 modules, and sorted them from largest to smallest by their sum of degrees. In different modules, the composition of their classes is different. It demonstrated that the fungal communities in the NSCLC were remodeled. The power-law distribution which means that most species have a small number of connections and very few species have a very large number of connections is a typical scale-free network that indicates that a fungal community has been constructed [27, 28]. Moreover, we re-visualized the network and focused on 4 main modules (Fig. 3a). The "dominant fungi" with highest relative abundance in the 4 modules were
*Dothideomycetes, Leotiomycetes and Eurotimycetes.* Interestingly, it was similar to the more connected fungi which had more degrees in Fig. 2a. Therefore, we had reasons to believe that the change in the abundance of fungi may be related to the degree of communication of different microbial flora.

Co-occurrence (positive association) may reflect the overlap of niches, while co-exclusion (negative association) may be caused by niche differentiation. From this point of view, we can infer the importance of non-biological environmental selection or competition between organisms in the construction of fungal communities. When phylogenetically closely related species tended to co-occur in the network, and distantly related species tended to co-exclude, it signified that the deterministic process which means environmental factors affect community structure occupied a major position in the construction of fungal communities. Conversely, when closely related species tended to co-exclude, and distantly related species tended to co-occur, the competitive exclusion may be dominant in community [29]. In this study, there were more co-occurrences in NSCLC and less co-exclusion. Moreover, the same fungal classes tended to co-occur. Therefore, we hypothesized that it may be the occurrence of NSCLC that led to changes in the living environment of the fungal community, indicating that environmental selection played a major role in the reorganization of the fungal community. Also the result in figure S3 (Additional file3) showed that the species that was the network hub was significantly distinct in NSCLC and non-cancer controls ($p = 0.0027$). In module 1, the relative abundance of *Dothideomycetes* was higher than half of the total abundance. It was therefore deduced that module 1 may be most affected by environmental factors and our following analysis confirmed that nearly all of the relative abundances of species detected as module1 hubs in NSCLC were significantly different from species in non-cancer controls.

In module 2 which had the most abundant overall connectivity of NSCLC, *Leotiomycetes* had the highest relative abundance, followed by *Eurotimycetes* and *Dothideomycetes*. *Malasseziomycetes* deserved our attention because it did not appear in module 2 but existed in the other three modules. Modules may reflect the heterogeneity of habitats, the aggregation of phylogenetic closely related species, the overlap of niches, and the co-evolution of species, which are phylogenetic, evolutionary, or functionally independent units [23]. We used each module to do regression analysis and it can be seen that except for module 2, the overall abundance of the other 3 modules increased with the progression of NSCLC and that module 1 was the best fit one. So according to Perez-Valera et's theory, module 1 represented fungi that were altered by changes in the airway environment.

According to the connectivity within the modules ($Z_i$) and between modules ($P_i$), all nodes were divided into peripherals, connectors, module hubs and network hubs. The last 3 nodes can be considered as the key nodes in entire network. Among the 3 categories, network hubs and module hubs were more worthy of attention. The former may be the core of the entire NSCLC network, and the latter was the core fungi of the corresponding modules. Among them, the network hub was *Pleurotus ostreatus*, and module hubs included five species which were *Vanderwaltozyma polyspora*, *Apiotrichum porosum*, *Lobosporangium transversal*, *Capronia epimyces* and *Pseudomicrostroma glucosiphilum*. We also found that abundance of the 3 super-generalists were significantly higher in the NSCLC group than in non-cancer controls (Fig. 3c). Therefore, changes in abundance may be the underlying cause of shift in the fungal community.
in NSCLC group. Finally, regarding the specific species in NSCLC, we found *Alternaria arborescens* was significantly different between two groups and its abundance was increased with cancer progression. More and more studies have shown that changes in the microbial flora may be related to the occurrence of cancer [30], but fewer studies have focused on fungi. This study innovatively proposes that changes in the structure of fungal communities in lung cancer may be closely related to changes in fungal abundance caused by cancer.

**Conclusions**

In our research, we found that with the progress of NSCLC, the diversity of fungi in the airway is increasing, and the community structure of fungi has changed compared with the control group. Some fungi that are more active in the community may play a key role. The relative abundance of *Alternaria arborescens* gradually increased with the progression of NSCLC, so it is speculated that this fungi may play some role in the occurrence and development of NSCLC. In this study, we innovatively studied the changes of fungi in the airways of patients with NSCLC from the perspective of fungal community structure. It provides new ideas for the treatment and diagnosis of NSCLC in the future.

**List Of Abbreviations**

| The phrase full name                  | abbreviations |
|--------------------------------------|---------------|
| non-small-cell lung cancer           | NSCLC         |
| bronchoalveolar lavage               | BAL           |

**Declarations**

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

All authors have agreed on the consent of the manuscript.

**Availability of data and material**

All the data generated or analyzed during this study are included in this published article and its supplementary files. The metagenomic sequencing data presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: SRA, PRJNA736814.

**Competing interests**
The authors declare that they have no competing interests.

**Funding**

This work was funded by the National Natural Science Foundation of China (81972198, 81773147, 32000054), the Natural Science Foundation of Hunan Province (2020JJ5702), Strategic Priority Research Program of Central South University (ZLXD2017004); Key Research and Development Program of Hunan (2019SK2253).

**Authors’ contributions**

JQT, ZY, JJX and JW designed the experiments, JQY carried out experiments, YMZ prepared the figures. YMZ and JQY drafted the manuscript. LYC, WZ, MHW participation in discussion and revised the manuscript. All authors contributed to this manuscript. All authors read and approved the final manuscript.

**Acknowledgements**

We thank Professor Jing Huang from department of Parasitology of Central South University for providing the modifications to this article.

**References**

1. Okazaki I, Ishikawa S, Ando W, Sohara Y. Lung Adenocarcinoma in Never Smokers: Problems of Primary Prevention from Aspects of Susceptible Genes and Carcinogens. Anticancer Res. 2016;36(12):6207-24.

2. Gosens R, Hiemstra PS, Adcock IM, Bracke KR, Dickson RP, Hansbro PM, Krauss-Etschmann S, Smits HH, Stassen FRM, Bartel S. Host-microbe cross-talk in the lung microenvironment: implications for understanding and treating chronic lung disease. Eur Respir J. 2020;56(2).

3. Ramirez-Labrada AG, Isla D, Artal A, Arias M, Rezusta A, Pardo J, Galvez EM. The Influence of Lung Microbiota on Lung Carcinogenesis, Immunity, and Immunotherapy. Trends Cancer. 2020;6(2):86–97.

4. Morris A, Beck JM, Schloss PD, Campbell TB, Crothers K, Curtis JL, Flores SC, Fontenot AP, Ghedin E, Huang L et al. Comparison of the Respiratory Microbiome in Healthy Nonsmokers and Smokers. Am J Resp Crit Care. 2013;187(10):1067-75.

5. Zheng L, Sun R, Zhu Y, Li Z, She X, Jian X, Yu F, Deng X, Sai B, Wang L et al. Lung microbiome alterations in NSCLC patients. Scientific reports. 2021;11(1):11736.

6. Willinger B. Emerging Fungi and Diagnosis of Fungal Infections: Current Knowledge and New Developments. J Fungi. 2021;7(4).

7. Cui LJ, Morris A, Ghedin E. The human mycobiome in health and disease. Genome Med. 2013;5.
8. Hamad I, Raoult D, Bittar F. Repertory of eukaryotes (eukaryome) in the human gastrointestinal tract: taxonomy and detection methods. Parasite Immunol. 2016;38(1):12–36.
9. Hamad I, Sokhna C, Raoult D, Bittar F. Molecular Detection of Eukaryotes in a Single Human Stool Sample from Senegal. Plos One. 2012;7(7).
10. Ghannoum MA, Jurevic RJ, Mukherjee PK, Cui F, Sikaroodi M, Naqvi A, Gillevet PM. Characterization of the Oral Fungal Microbiome (Mycobiome) in Healthy Individuals. Plos Pathog. 2010;6(1).
11. Gelders YG, Reytjens AJ, Ash CW, Aerts TJ. 12-month study of haloperidol decanoate in chronic schizophrenic patients. Int Pharmacopsychiatry. 1982;17(4):247–54.
12. Chen SF, Zhou YQ, Chen YR, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):884 – 90.
13. Xu HB, Luo X, Qian J, Pang XH, Song JY, Qian GR, Chen JH, Chen SL. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. Plos One. 2012;7(12).
14. Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. Biodata Min. 2015;8.
15. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3).
16. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010;26(11):1463-4.
17. Thukral AK. A review on measurement of Alpha diversity in biology. Agricultural Research Journal. 2017;54(1).
18. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003;14(6):927 – 30.
19. Watts SC, Ritchie SC, Inouye M, Holt KE. FastSpar: rapid and scalable correlation estimation for compositional data. Bioinformatics. 2019;35(6):1064-6.
20. Newman MEJ. Modularity and community structure in networks. P Natl Acad Sci USA. 2006;103(23):8577-82.
21. Deng Y, Jiang YH, Yang YF, He ZL, Luo F, Zhou JZ. Molecular ecological network analyses. Bmc Bioinformatics. 2012;13.
22. Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. Nature. 2005;433(7028):895–900.
23. Olesen JM, Bascompte J, Dupont YL, Jordano P. The modularity of pollination networks. P Natl Acad Sci USA. 2007;104(50):19891-6.
24. Zhou JZ, Deng Y, Luo F, He ZL, Tu QC, Zhi XY. Functional Molecular Ecological Networks. Mbio. 2010;1(4).
25. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10).
26. Brandon-Mong GJ, Shaw GTW, Chen WH, Chen CC, Wang DR. A network approach to investigating the key microbes and stability of gut microbial communities in a mouse neuropathic pain model.
27. Barberan A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence patterns in soil microbial communities. Isme J. 2012;6(2):343–51.

28. Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY, Chow CET, Sachdeva R, Jones AC, Schwalbach MS et al. Marine bacterial, archaean and protistan association networks reveal ecological linkages. Isme J. 2011;5(9):1414-25.

29. Perez-Valera E, Goberna M, Faust K, Raes J, Garcia C, Verdu M. Fire modifies the phylogenetic structure of soil bacterial co-occurrence networks. Environ Microbiol. 2017;19(1):317–27.

30. Borchmann S. An atlas of the tissue and blood metagenome in cancer reveals novel links between bacteria, viruses and cancer. Microbiome. 2021;9(1).

Figures
Figure 1

Lung microbiome characterization and fungal diversity. a. A Stack Graph which contains the top 15 species ordered by relative abundance. b. The \( \alpha \) diversity across different groups (NSCLC and non-cancer controls). Violin graphs in blue or orange show the Shannon-Wiener index and Gini-Simpson index of NSCLC and non-cancer controls, respectively, at species level (Wilcox test, \( p < 0.05 \)). c. PCoA based on the Bray-Curtis dissimilarity index shows the between-subjects \( \beta \) diversity across groups, in which the
blue circles and orange triangles represent NSCLC and non-cancer controls, respectively (PERMANOVA, p = 0.0012).

Figure 2

a. The microbial association networks where different circles indicate different fungal classes and pie charts represent the composition of classes between NSCLC and non-cancer controls. Red links indicate positive covariation between two individual nodes, whereas blue links indicate negative covariation. b.
The sum of degrees in two groups. c. The boxplot that shows the different degree in network of non-cancer controls and NSCLC patients (Wilcox test, p < 0.0001). d. A rose graph to show degrees in different tumor module. e. f. g. The relationship between different $\alpha$ diversity indices and cancer stage. The lines denote the least-squares linear regressions across cancer stage, with their 95% confidence intervals (gray-shaded areas).

Figure 3
a. Network diagram with nodes colored according to each of the four main ecological clusters (modules 1–4). Each node is the different species in the cancer group and different colored dots represent different modules, and the sizes of different dots represent different degrees. 
b. Relative abundance properties of the different fungal classes in the four main ecological clusters. 
c. The relationship between different relative abundance of modules and cancer stage. 
d. A graph to show each node in a. with four types (peripherals, connectors, module hubs, network hubs) classified by within module connectivity and among module connectivity.
a. Volcano plot demonstrates the differential abundance of fungal between NSCLC and non-cancer controls. A Wald test was used to calculate the p values. Points are colored according to the number of log2 fold change if they passed significance thresholds (P < 0.05) c. Venn graph shows the number of different fungi or hub species filtered by different methods. b. The variation between the relative abundance of the fungi in all three results in b. and cancer stage. b. A heatmap graph to show the relative abundances of the 44 biomarker fungi species predicted by Random Forest and DESeq2. The heatmap is divided into 3 parts with different colors. The graph colored with orange, blue and green is the results of Random forest, DESeq2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx
- Additionalfile3.docx
- Additionalfile4.docx