TCGA Microbiome Analysis

Eeman Abbasi [1], **Erol Akçay** [1]
[1] Department of Biology, University of Pennsylvania
433 S University Ave, Philadelphia, PA 19104, USA

May 10, 2023

1 Introduction

The microbiome refers to a diverse collection of microorganisms that inhabit a host, including bacteria, viruses, fungi, and archaea. This ecological community resides within a dynamic host environment and is not isolated, but rather influenced by its social environment that is composed of other microbes. Microbes interact with each other through ecologically-mediated interactions such as competition for resources, cooperative or syntrophic interactions where they provide metabolites to each other, commensal, exploitative and amensal interactions. These interactions play a critical in determining the ecological characteristics of the microbial community. They inform which species can colonize, their abundances and how the community responds to the introduction of foreign microbial species. Understanding these complex interactions is critical to understanding the role of the microbiome in health and disease. However, in addition to the within-species interactions, the host itself can impose constraints on the microbial community through diet and immune response. Host immune control can act as an instance of leash on the microbial community, where the immune response can dictate the community membership and the size of the microbiome. Recent literature suggests that the host's immune control plays a critical role in shaping the ecological characteristics of the microbiome. Changes in the host's immune state have been linked to a dysbiotic configuration of the microbial community, where deviation from the homeostatic immune state is linked to changes in microbial diversity and abundance.

Host immune control and between species interactions in union are believed to determine the microbiome community structure. Host immune control, when modeled theoretically to globally regulate microbial community abundance, can have differential effects on the microbial community based on the prevalent ecological interaction type in the community. In

instances of inflammation-prone immune phenotype, highly mutualistic communities will be most affected by the host immune control. This is because species engaged in cooperative interactions build each others population sizes through the mutual exchange of essential growth promoting resources. Inflammation-prone phenotype is indicative of a tighter leash by the host on the microbiome. This results in greater selection against high microbial community sizes, that selects for communities that are inturn more competitive in nature as such communities appear resilient to the changes in host immune control. Competitive communities remain independent of the other species abundances as they rely on the external supply of resources and not on species in the community, and species within this community lower population sizes of other species through competitive exclusion. The shift towards more competitive communities at an inflammation-prone phenotype has consequences on the resulting species diversity and abundances observed in the microbiome (Figure 1). Whereas in the immune-dessert state of the host, mutualistic communities will be able to establish themselves and thrive, consequently allowing for greater species diversity and abundance (Figure 1). These theoretical insights need to be validated using real-world microbiome data to further solidify our understanding of the the impact of host immune control on the microbial community.

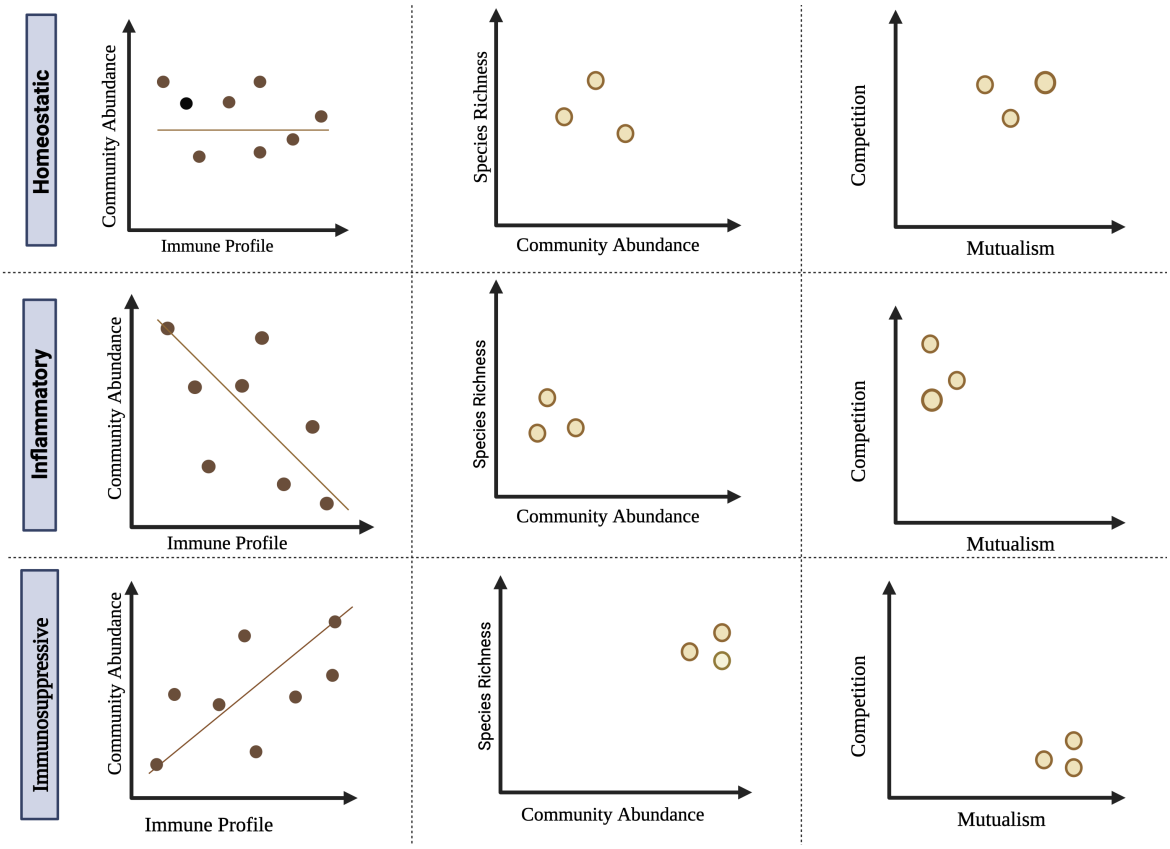


Figure 1: Theoretically driven insights on the role of varying host immune states on the microbiome community structure.

The Cancer Genome Atlas (TCGA) is a valuable resource for this purpose, offering comprehensive information on the genomic, transcriptomic, proteomic, microbiome, and immune profiles of 33 different cancer types. Exploring microbial ecological patterns associated with cancer is particularly useful given that cancer represents a perturbed state of the host, characterized by disrupted immune response. Each cancer type has a unique tumor microenvironment (TME) comprising varying amounts of immune populations. Certain cancer types are categorized as immunologically "quiet" or "cold" tumors due to their limited immune cell infiltration, such as Breast, Prostate, Glioblastoma, and Pancreatic cancer. Other cancers, such as Melanoma, non-small cell lung cancer, head and neck, and Liver cancer, are labeled as "hot" tumors with immune-inflamed phenotype because they exhibit high levels of immune cell infiltration and inflammation. There are also cancer types that fall in between these two immune phenotypes. Thus each distinct immunophenotype of a cancer will consequently allow for a distinct microbiome community. Hence uncovering how the microbial community differs across a range of cancer types, and hence across an immune gradient provides us with an opportunity to specifically answer what is the role of host immune control on the microbiome community.

2 Methods

2.1 Microbiome Dataset

We downloaded the cancer microbiome dataset which is publically available at: https://github.com/knightlab-analyses/mycobiome. The dataset comprises fungal and bacteria species analysis conducted for over 17,000 patient tissue, blood, and plasma samples across 35 cancer types. The dataset comprises an operational taxonomic unit table with information available at the species order level. The dataset has already been corrected for any batch effects and contamination. Immune profiles were accessed from Thorrson et al. 2018.

2.2 Determining co-occurring communities:

2.2.1 Fast greedy approach

For each cancer type we constructed a correlation matrix, calculating a pairwise coefficient for each species in the community. The correlation coefficient is determined using microbial species abundances, where a positive value indicates that the species abundances vary in the same direction suggesting an underlying common process influencing both species abundances. Whereas a negative value is indicative of both species' abundances varying in the opposite directions. We then used this correlation matrix as an input to a community detection algorithm to determine co-occurring communities within the sample. We used the fast greedy algorithm to merge species into co-occurring groups. The algorithm treats each species as a separate mode and then iteratively joins them into clusters based on the modularity of each node in the network. It merges pairwise nodes such that the union further increases the modularity in the network. This process gets repeated until no further increase in modularity can be achieved.

2.2.2 Non Matrix Factorization Approach

Currently statistical techniques (NMF) developed for the discovery of mutational signatures in the genome are now being applied to uncover microbial signatures associated with the host state. An original matrix is factorized into two matrices such that when they are multiplied together they return the original matrix. It provides a consortia of microbes that are more likely (provides probability for how likely we will observe a given microbe in a community) to be present in a community based on their abundance levels observed in the dataset. The

algorithm can provide a unique clustering of the microbial communities such that a species belonging to a particular group can also be a resident of another community. We used the NMF algorithm to find microbial communities that are more likely to co-occur than by random chance.

2.3 Microbial Richness and Abundance

After determining the co-occurring communities for a specific cancer type, we then determined the number of observed counts of species within a co-occurring group and the overall group abundance from the the OTU table.

2.4 Competition and cooperation landscape

2.4.1 Construction of a species and a community level metabolic model

Genome-scale metabolic model was reconstructed using a python package CarvMe (Machado et al., 2018). The CarvMe package allows fast automated reconstruction of genome-scale metabolic models for microbial species. The framework, unlike other metabolic reconstruction tools, uses a top-down approach constructing a universal metabolic model that is later carved to produce an organism specific metabolic model. The carving process involves the removal of any reactions or metabolites that are not predicted to be present for that microbial species based on the species genetic sequence. Individual species genome-scale metabolic models can be merged to create a community level metabolic model, preserving the core metabolism specific to each of the microbial species. We provided refseq accession NCBI code when using the CarveMe to download species-specific genomes in the co-occurring groups to build metabolic models (xml format).

2.4.2 Determine microbial metabolic resource requirements

The metabolic models can then be expressed mathematically as a stoichiometric matrix (S) whose rows represent the metabolites and the columns represent the reactions taking place in the metabolic network. We can then use this S matrix as an input to the flux balance analysis. Flux balance models for several organisms have been used to define the range of theoretically possible phenotypes for a given genotype by simulating environmental conditions. Using flux balance analysis we can determine the fewest number of metabolites exchanged among

organisms under the constraint that species will be able to produce a biomass greater than a threshold value.

We use the SMETANA algorithm developed by Zelezniak et al. (2015) to search for minimal media required by a microbe based on its stoichiometry. SMETANA is an extension of the algorithm developed by (Klitgord and Segr'e, 2010) which uses the flux balance analysis to search for interaction inducing media. The algorithm requires a species or community metabolic model as an input and can determine a set of metabolites or the minimal media that is needed by the community for growth. Removal of any single metabolite from the minimal media will render the community unfeasible. The algorithm can be applied in an:

- Non-interacting environment where species can only acquire resources from their abiotic environment.
- Interacting environment where species can acquire resources from both the external environment and also the metabolites secreted from microbial species present in the community.

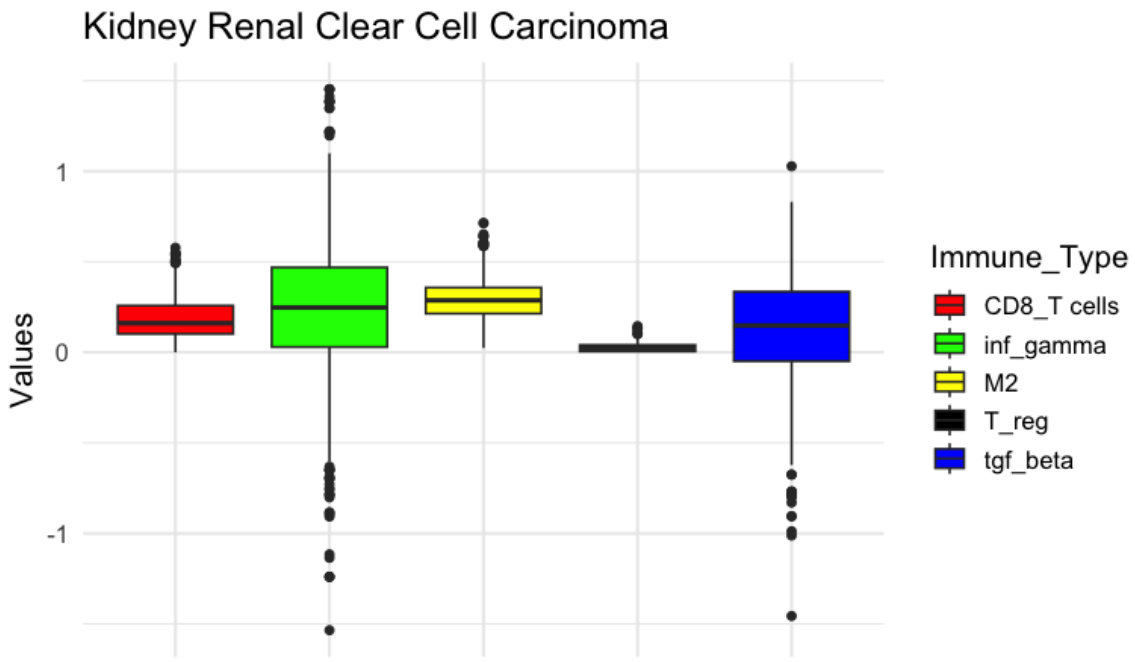
The algorithm assigns an initial minimal media consisting of a set of metabolites that are needed by either of the species present in the community based on their stoichiometric profiles. The initial media is then perturbed, where all metabolites that can provide multiple elements such as amino acids being a source of both carbon and nitrogen are substituted. A Flux balance analysis minimization is performed on the each metabolite and is repeated iteratively to find a alternative metabolite that restores the capacity for growth for the community. The microbes are then simulated to grow as a community and individually in the putative minimal media. If microbes can only grow in a community context in the chosen media and not individually then that particular media induces obligate syntrophic interactions.

Using this framework we can determine the minimal resource requirements of a microbe, and of the microbial community. In addition to that, determining the minimal media in a non-interacting and an interacting environment can help us identify the metabolites acquired from the abiotic environment versus the metabolites that are exchanged among community members.

Competition: Metabolic resource overlap between species in the community
Mutualism: Minimal resource requirement in an non-interacting environment - minimal resource requirement in a interacting environment.

2.5 Immune state and its relation to the microbiome

We currently have access to a range of immune cell gene expression counts across a range of immune cell types across all cancer types including:



2.6 Preliminary Results

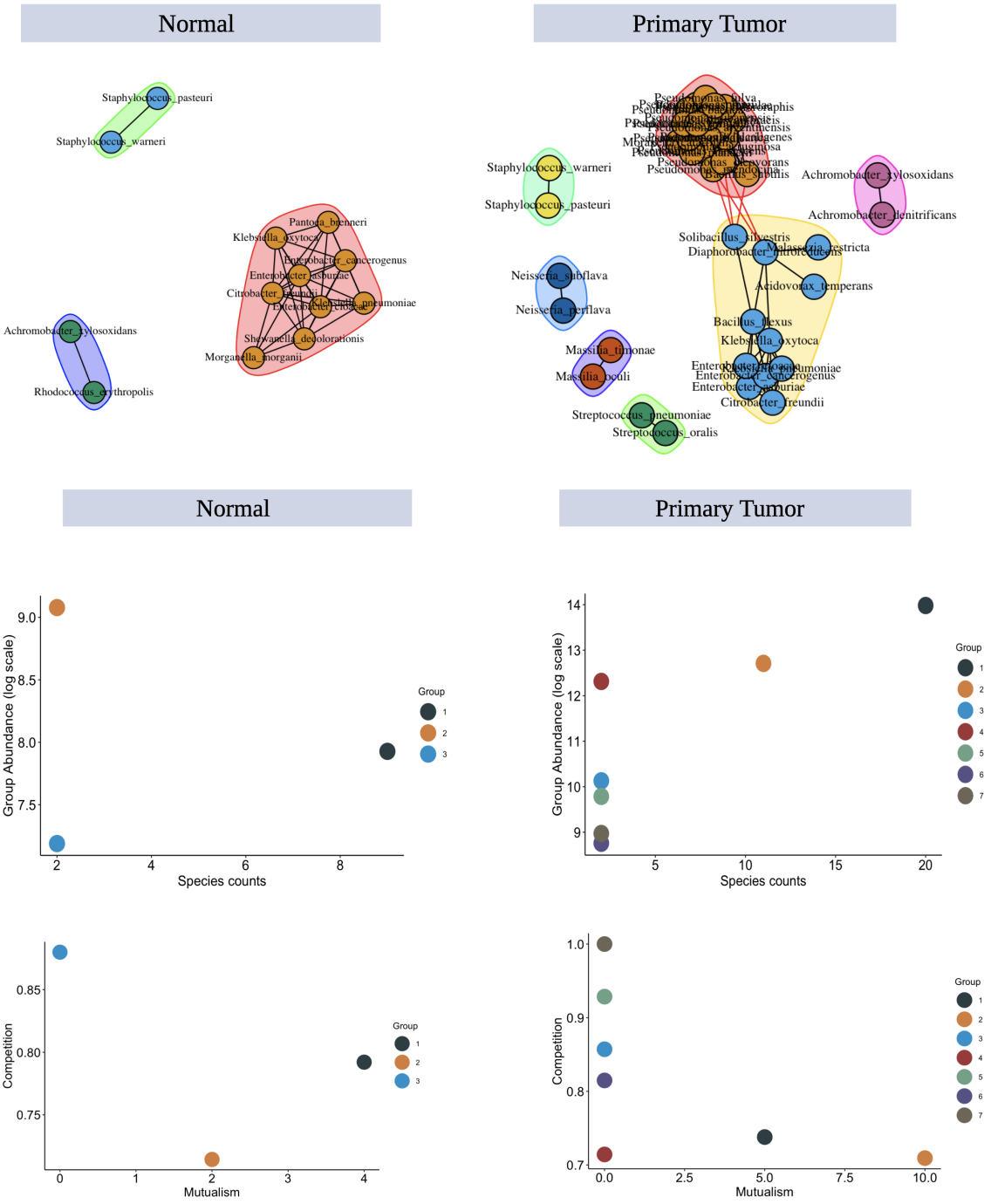


Figure 2: Represents microbial community characteristics for ovarian cancer. A network is shown, representing the co-occurring communities formed in both normal and primary tumor samples using a fast-greedy algorithm, and two scatter plots i) microbial co-occurring group abundance and species counts, and ii) Competition and mutualism score calculated for each co-occurring community.

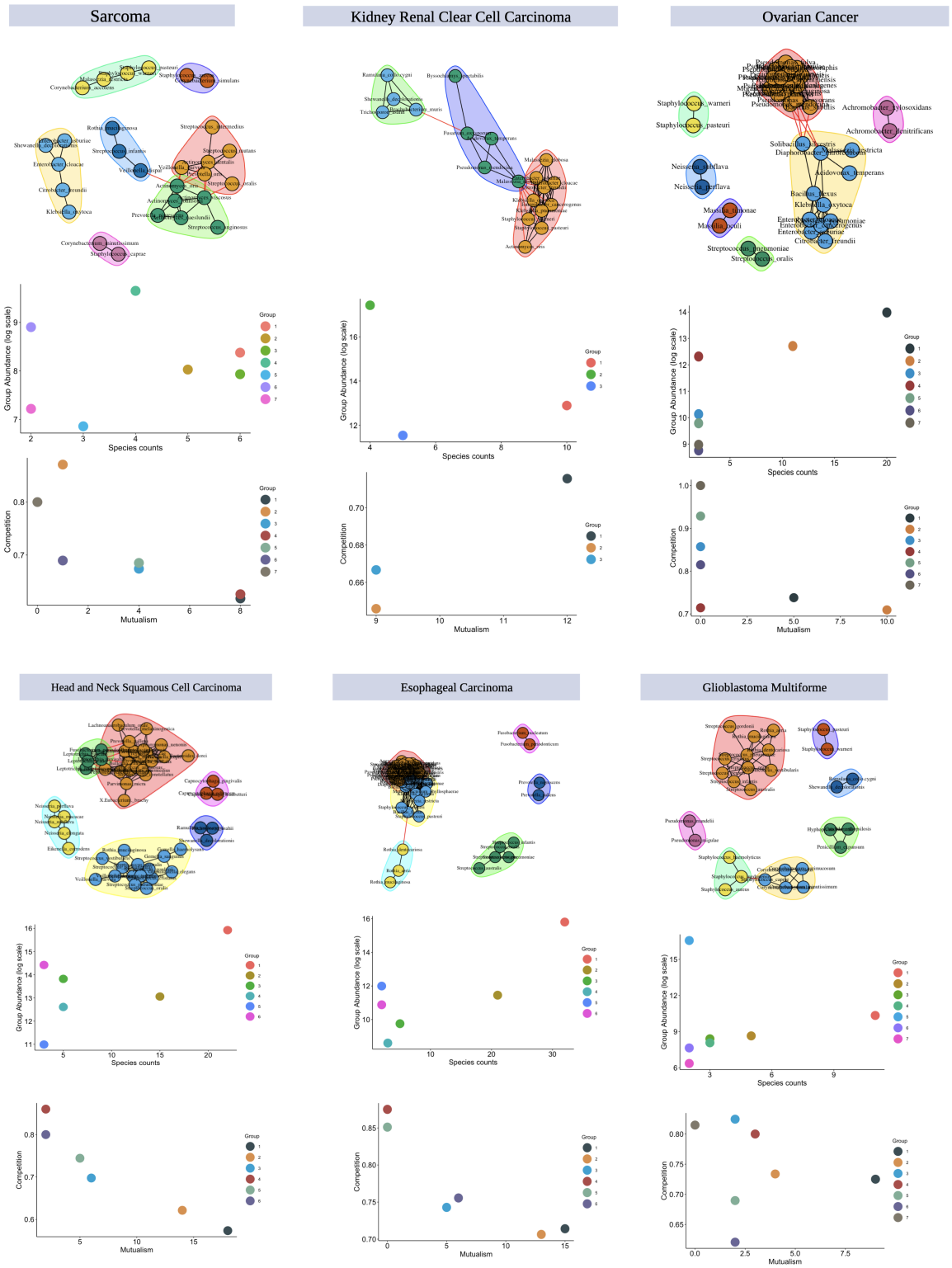


Figure 3: Represents microbial community characteristics across a range of cancer types. For each cancer type, there is a network representing the co-occurring communities formed using a fast-greedy algorithm, and two scatter plots i) microbial co-occurring group abundance and species counts, and ii) Competition and mutualism score calculated for each co-occurring community.

2.7 Future Directions

1. Conduct a PCoA of immune variables and determine the main axis of variation that explains the inflammation and immune-dessert phenotype. Correlate the axis variables with mutualistic or competitive co-occuring communities present across a range of cancer types.
2. Rerun the co-occurring community analysis using NMF (if needed) - compare the fast greedy approach and the NMF.