Preliminary concepts for taxonomic aggregation using modified isometric log ratio transformation (ILR)

Quang Nguyen

Dartmouth College

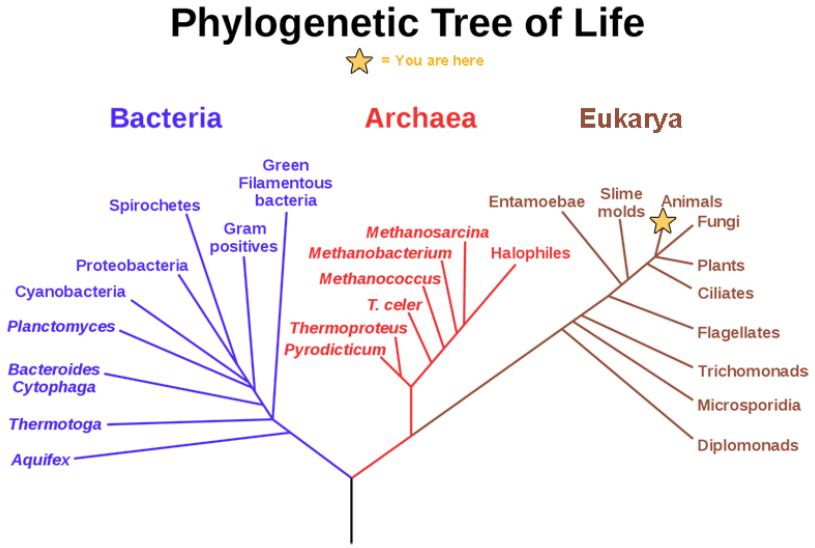
May 3, 2020

Microbiome data is strictly compositional

- All short read sequencing data exists as relative abundances
- RNASeq-type data sets usually can rely on certain assumptions about the relative consistent expression of housekeeping genes (e.g. DESeq2) or specific measurements of the PCR process (UMI counts in scRNAseq)
- Analysis of microbiome data hence relies on log-ratio based analyses from the compositional data analysis literature.

Microbiome analysis involves a lot of aggregation

- Most often, variables are aggregated to the Linnean taxonomic tables
- **Note:** Phylogeny and taxonomy generally agrees, however this is not always the case.
- Aggregating variables to pre-defined sets describes a class of methods known as gene set analysis (GSA).



Typical analysis involves summation of relative abundances (or raw counts)

- This approach might cause problems
  - The compositional center is changed. For example, take a simple composition of 3 parts $x = [x_1, x_2, x_3]$ and the aggregated composition $y = [x_1 + x_2, x_3]$ (with n samples). The center [**?**] of the initial composition is

$$cen(x) = \mathcal{C}\left[(\prod^{n} x_1)^{1/n}, (\prod^{n} x_2)^{1/n}, (\prod^{n} x_3)^{1/n}\right]$$

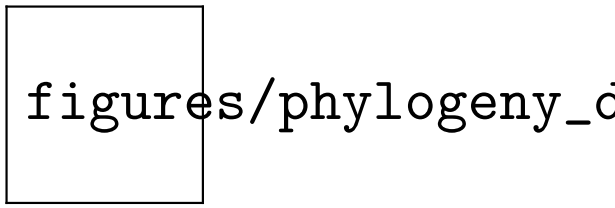
    while the center of the aggregated composition is

$$cen(y) = \mathcal{C}\left[(\prod^{n} (x_1 + x_2))^{1/n}, (\prod^{n} x_3)^{1/n}\right]$$

  - This results in the lack of preservation for inter-sample distances following aggregation, and the relationship is non-monotonic [**?**].

Isometric log-ratio transformation

- Defined as a transformation of of the composition in the simplex S^D to $\mathcal{R}^{D-1}$ by projecting onto an Aichison orthonormal basis.
- Prior researchers have proved [**?**] that such a basis can be constructed from a sequential binary partition (SBP) or a tree - each variable is therefore defined as a "balance" [**?**]



Isometric log-ratio transformation

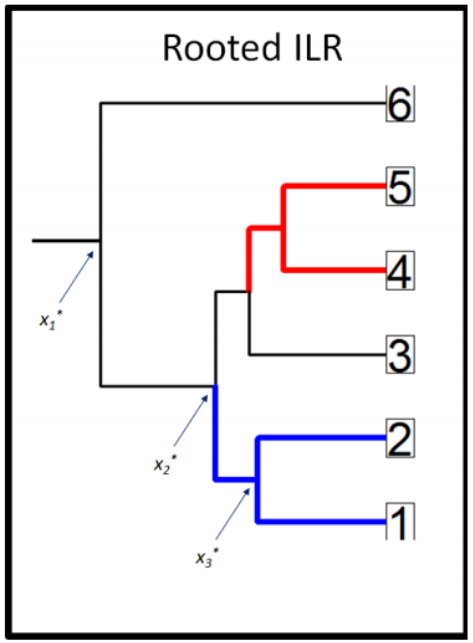
- ILR statistic is defined as

$$x^*_{R/S} = \sqrt{\frac{rs}{r+s}} \log(\frac{g(\boldsymbol{y}_R)}{g(\boldsymbol{y}_S)} \qquad (1)$$

- $g(\boldsymbol{y}_R)$ indicates the geometric mean of all y_i for $i \in R$
- The basis defined by the R/S partition is

$$\boldsymbol{e}_i = \mathcal{C}[\exp(0, .., \underbrace{a, ..., a}_{r \text{ elements}}, \overbrace{b, ..., b}^{s \text{ elements}}, ..., 0))]$$

where $a = \sqrt{\frac{r}{(r+s)\cdot s}}$ and $b = \sqrt{\frac{-s}{(r+s)\cdot r}}$



Competitive gene set testing using isometric log ratio transformation

- $\boldsymbol{X}$: The $n \times p$ matrix of relative abundances of p LT proportions in n samples
- $\boldsymbol{A}$: The $m \times p$ matrix denoting the assignment of p LT proportions into m HT sets

We generate the following matrix:

- $\boldsymbol{S}$: The $n \times m$ matrix denoting the enrichment scores of m HT sets by n samples.

Competitive gene set testing using isometric log ratio transformation

With inputs $\boldsymbol{X}$ and $\boldsymbol{A}$, we compute $\boldsymbol{S}$ as follows:

1. Let $\boldsymbol{R}$ be a $n \times m$ matrix of competitive *ilr* statistic for each HT set which is defined as follows:

$$R_{ij} = \sqrt{\frac{\sum_k A_{jk}(p - \sum_k A_{jk})}{p}} \log\left(\frac{g(\boldsymbol{x}_{ik}|A_{jk} \neq 0)}{g(\boldsymbol{x}_{ik}|A_{jk} = 0)}\right) \tag{2}$$

2. To capture the distribution of the *ilr* test statistic under the null hypothesis that the relative weights of LT proportions in the HT sets to the overall composition is no different than LT proportions not in the HT sets, the competitive *ilr* statistic is computed for each HT set with permuting row labels of matrix $\boldsymbol{A}$, simulating random assignments of LTs into HT sets. Denote $\boldsymbol{A}_{perm}$ be the row-permuted version of $\boldsymbol{A}$ and $\boldsymbol{R}_{perm}$ be row-permuted version of $\boldsymbol{R}$

Competitive gene set testing using isometric log ratio transformation

1. We fit a gaussian distribution using the method of maximum likelihood for each column $\boldsymbol{R}_{perm}$. This stem from previous research treating the *ilr* coordinates as normally distributed [**?**].
2. Calculate a z-score using the fitted null distribution to compute HT specific enrichment scores. We formulate the target matrix S as z-scores for the normal distribution fitted on columns of $\boldsymbol{R}_{perm}$:

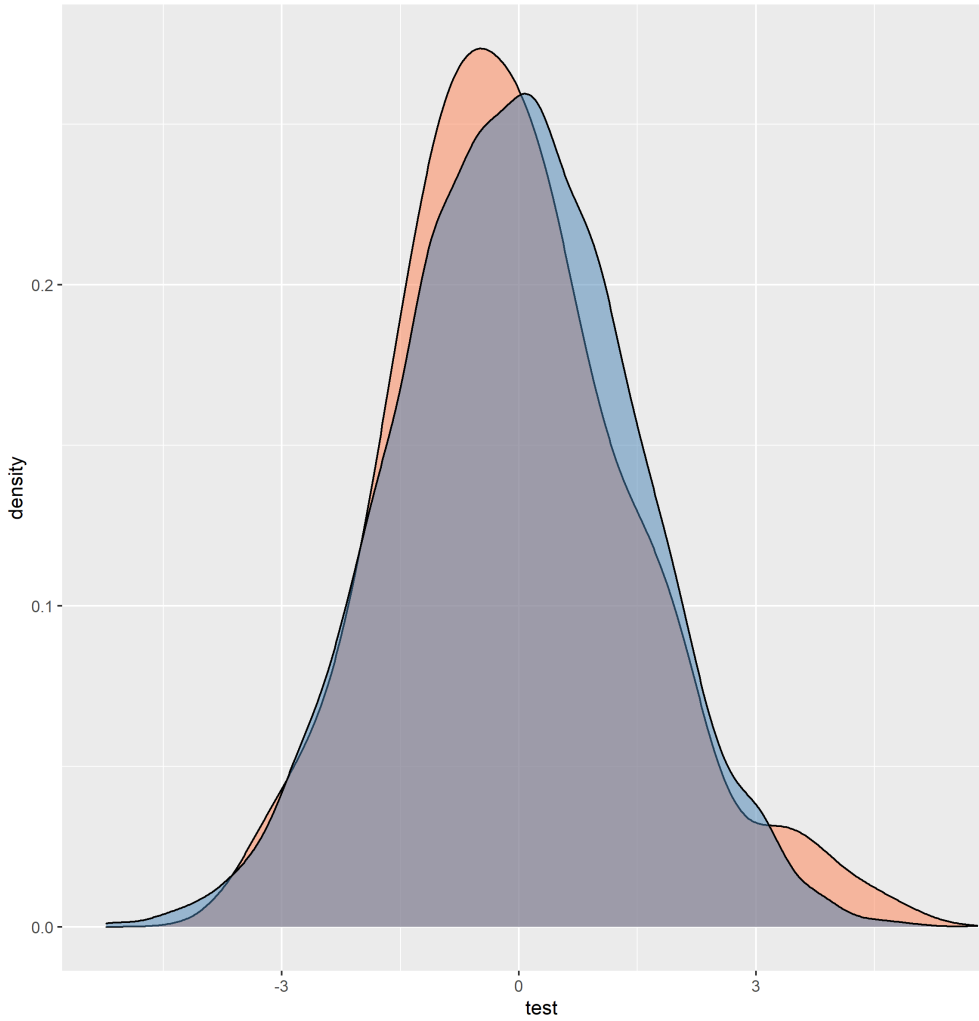
$$\boldsymbol{S}[, k] = (\boldsymbol{R}[, k] - \hat{\mu}_{\mathcal{N}_{\boldsymbol{R}_{perm}}})/\hat{\sigma}_{\mathcal{N}_{\boldsymbol{R}_{perm}}} \tag{3}$$

Statistical interpretation of this statistic

- Projecting the original composition onto the axis defined by the balance between taxa within the higher taxonomic set and those who are not.
- Equivalent to defining an *ilr* transform on a set of SBPs such that the first order split is between members of the set and the remaining variables.
- These coordinates are on the same scale, and is no longer bound by the composition. The variable aggregation process is also compositionally coherent.

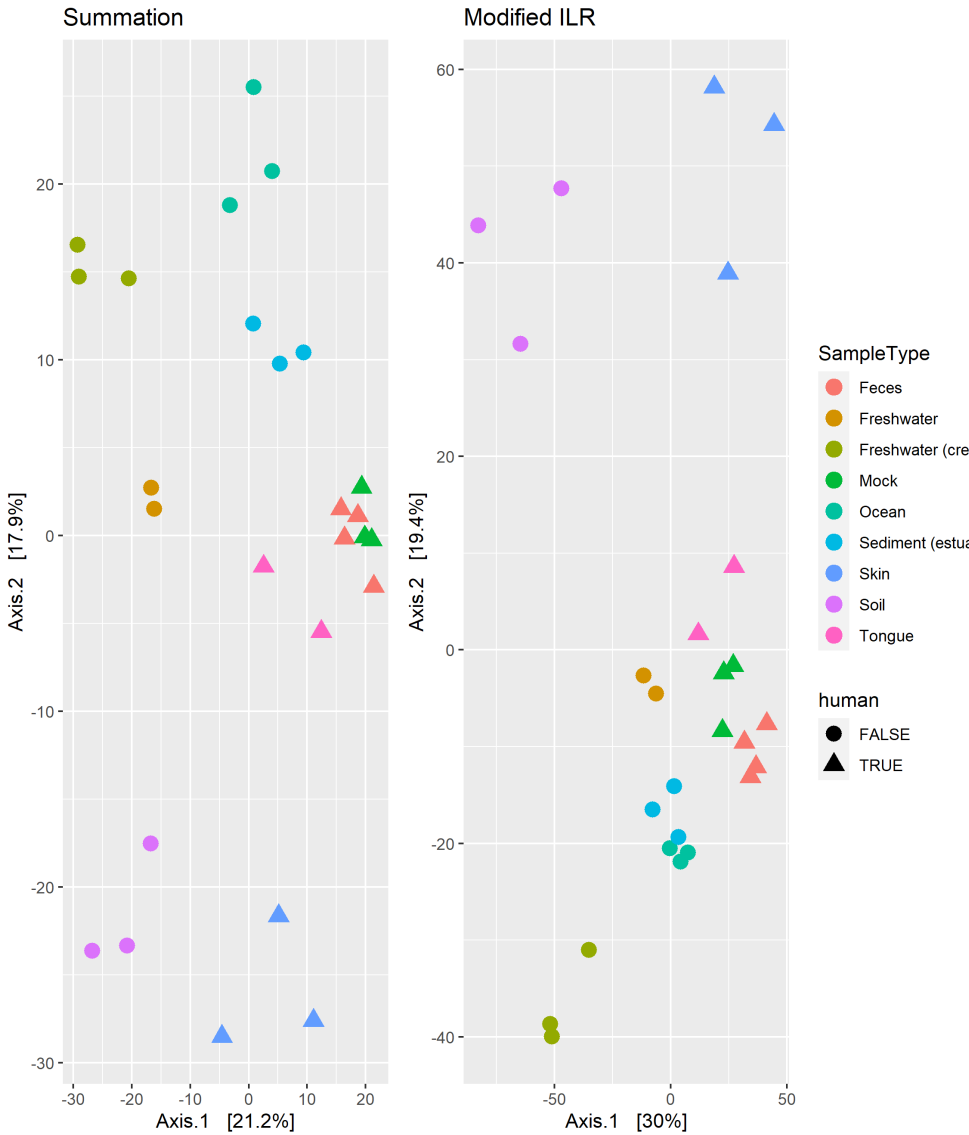
Early Results

- Density estimation for the null with 100 column permutations
- Mean: -0.07, SD: 1.5



Early Results

- GlobalPatterns data set.
- Genus level aggregation
- Euclidean distances (after *clr* transformation for simple summation counts)



Disadvantages

- Does not deal with true hierarchical structure
- Did not take into account phylogeny
- Unsure about justification above using simple sum
- Some uncertainties about correlation

Next Steps

- Implement the resampling procedure in the prototype
- Dealing with singletons and inter-taxa correlation
- Performing simulations to determine type I error and power.
- Real data implementation

Bibliography