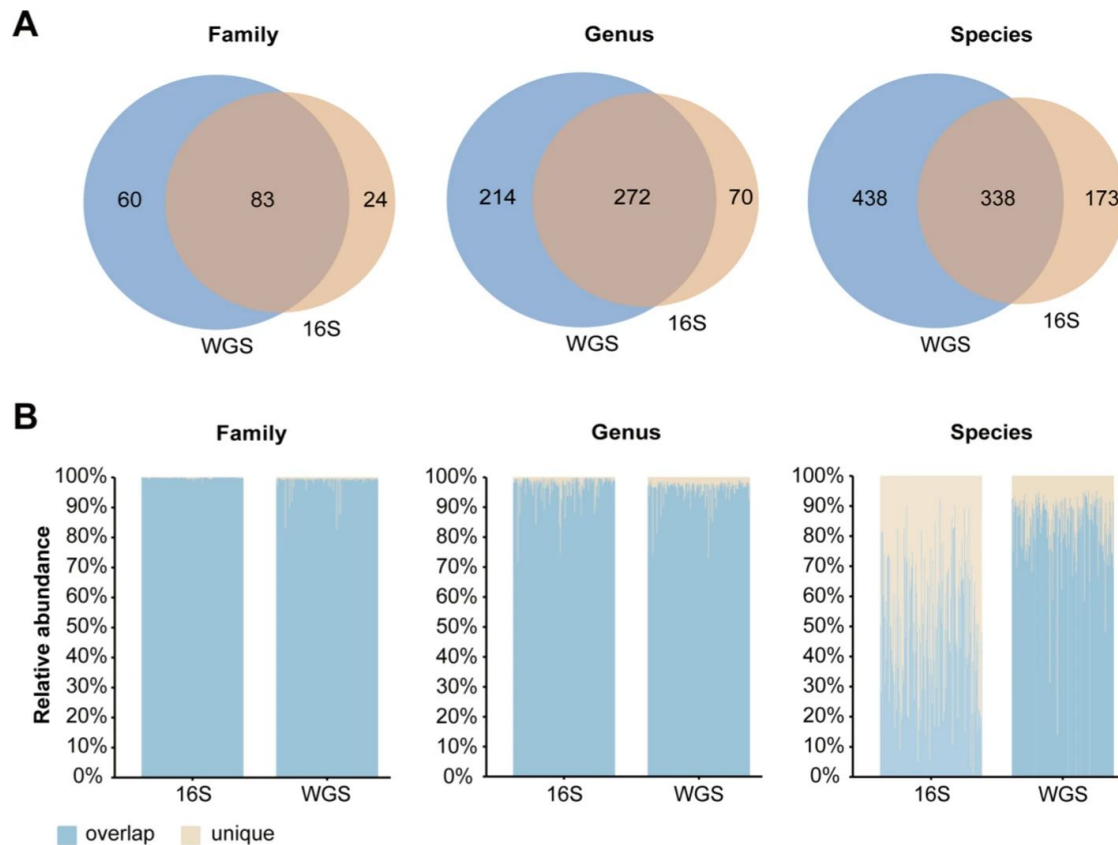


Pairing WGS and 16S data for enhanced representations: a proposal

Kevin Chen, November 2025

16S and WGS

- Given a sample, there are two different ways to figure out the bacterial composition
- **16S (cheap but low resolution)** sequencing reads the number of 16S sequences once amplified
 - one bacteria can have multiple 16S reads. Usually this is dependent on bacterial species/strain
 - The 16S sequences of some bacteria are more sensitive to being amplified/read
 - Issue: no species resolution
- **WGS (expensive)** (whole genome sequencing)
 - read and reconstruct the entire bacterial genome to determine whether a bacteria is present
- Currently, we have been learning with 16S sequence abundance counts



Differences between 16S and WGS at multiple taxonomic levels of in-house dataset 1. A Number of overlapped and unique taxa between 16S and WGS samples. B Richness of overlapped and unique taxa between 16S and WGS samples. <https://doi.org/10.1186/s12859-025-06156-7>

Why new approach

- “[The 16S gene] alone is not a reliable indicator of the functional ability of a microbiome.”¹
- “16S data may underestimate functional ability of microbiomes, due to strain-level adaptations to local conditions, or overestimate functional ability, due to functional redundancy across taxa.”¹
- The alternative is Whole Genome Sequencing (WGS)
- New masters thesis appears to show some success applying scGPT (the method we are using) to a smaller dataset (size 13,524) of WGS abundances.²
 - in contrast, we have been using a dataset of only 16S abundances (size 160,000)
 - They have more success with disease-prediction related tasks, which we are struggling with.

1. <https://doi.org/10.3389/fmicb.2020.00101>

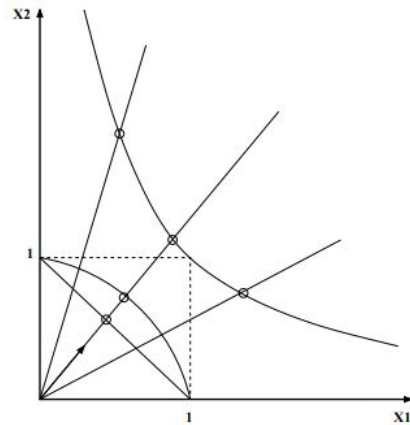
2. <https://dspace.mit.edu/bitstream/handle/1721.1/162973/medearis-medearis-meng-eecs-2025-thesis.pdf?sequence=1&isAllowed=y>

Compositionality

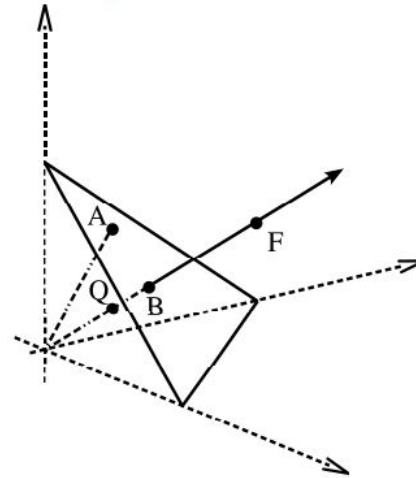
- Both are *compositional*, in this context meaning each datapoint represents ratios and therefore sum to 1
- Mathematically, **the sample space is a simplex**
- A simplex is a finite-dimensional inner-product space (called an Aitchison geometry)
- To convert to real number space, can transform with log operations
 - centre-log and isometric-log ratios are commonly chosen
- However, the data is sparse so this must be dealt with first (through imputation, zero-inflation, or other methods)

Compositionality

definition: parts of some whole which carry only **relative information** \iff compositional data are **equivalence classes**



compositional data in \mathbb{R}^2

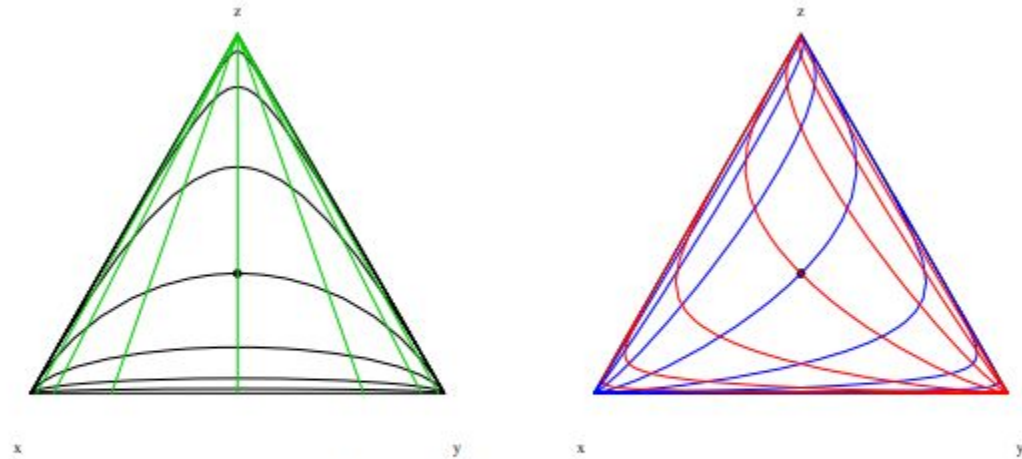


compositional data in \mathbb{R}^3

usual representation: subject to a **constant sum constraint**

V. Pawłowsky-Glahn
and
J. J. Egozcue

Compositionality



orthogonal grids in S^3 , equally spaced, 1 unit in Aitchison distance; the right grid is rotated 45° with respect to the left grid

Centre log ratio

Center log ratio transform [\[edit\]](#)

The center log ratio (clr) transform is both an isomorphism and an isometry where $\text{clr} : S^D \rightarrow U$, $U \subset \mathbb{R}^D$

$$\text{clr}(x) = \left[\log \frac{x_1}{g(x)}, \dots, \log \frac{x_D}{g(x)} \right]$$

Where $g(x)$ is the [geometric mean](#) of x . The inverse of this function is also known as the [softmax function](#).

Assumptions

- For a given genus, can we assume the given difference between 16S and WGS counts is a constant scaling (assuming sensitivity of 16S and WGS are constant but different for different bacteria)?
- In this case, in log space (after a centre-log-ratio or isometric-log-ratio transformation), the counts would be shifted by a constant vector.
- There may be cross-effects. Could be captured by linear model.

Potential Issues with these assumptions

- will have to do more literature search. Linearity may not be a good assumption
- what should we do if this is not a good assumption?

Models

- linear model
 - bayesian hierarchical model
 - training set: paired 16S and WGS representations from same sample
-
- An issue is that pairs of 16S and WGS is not commonly done.
 - dataset size probably only about 400.

Pipeline

- After converting existing 16S data to WGS data, rerun model
- Or, somehow integrate this transformation into the model