Composition statistics

What is compositional data?

Formally, x is a composition if

$$\sum_{i=0}^{D} x_i = c$$

and

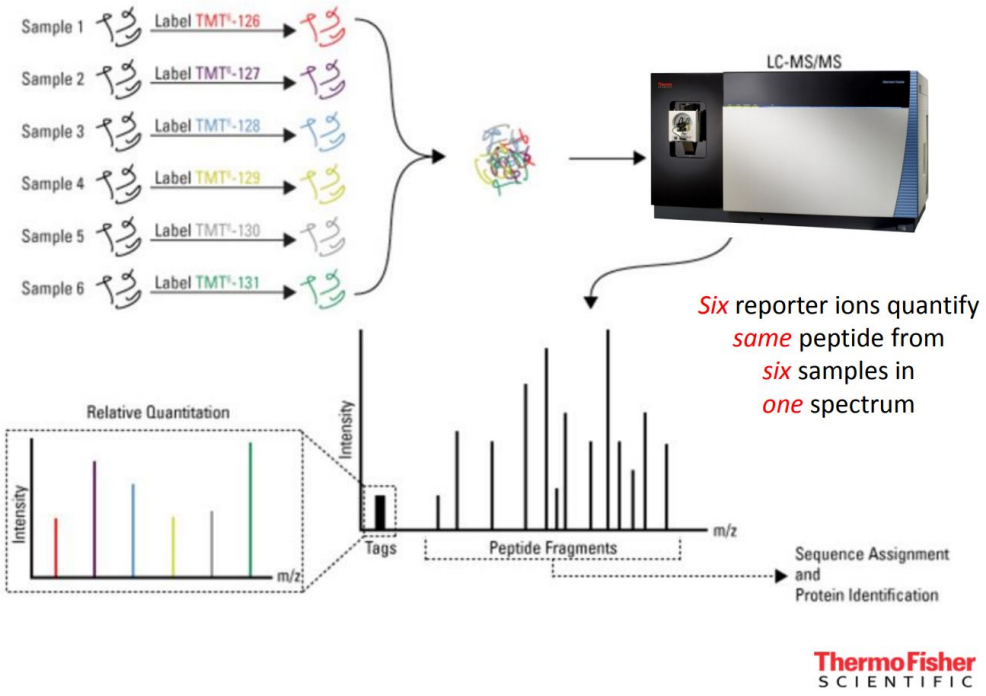
$$x_i > 0, 1 \leq i \leq D$$

and c is a real valued constant and there are D components for each composition. Commonly in this module c=1. Compositional data can be analyzed using Aitchison geometry.

However, in this framework, **standard real Euclidean operations such as addition and multiplication no longer apply**. Only operations such as perturboation and power can be used to manipulate this data (see wikipedia article about Compositional data.

(I.e. compositional data = **VALUES ARE RELATIVE**)

TMT-labeling is compositional data.



ThermoFisher SCIENTIFIC

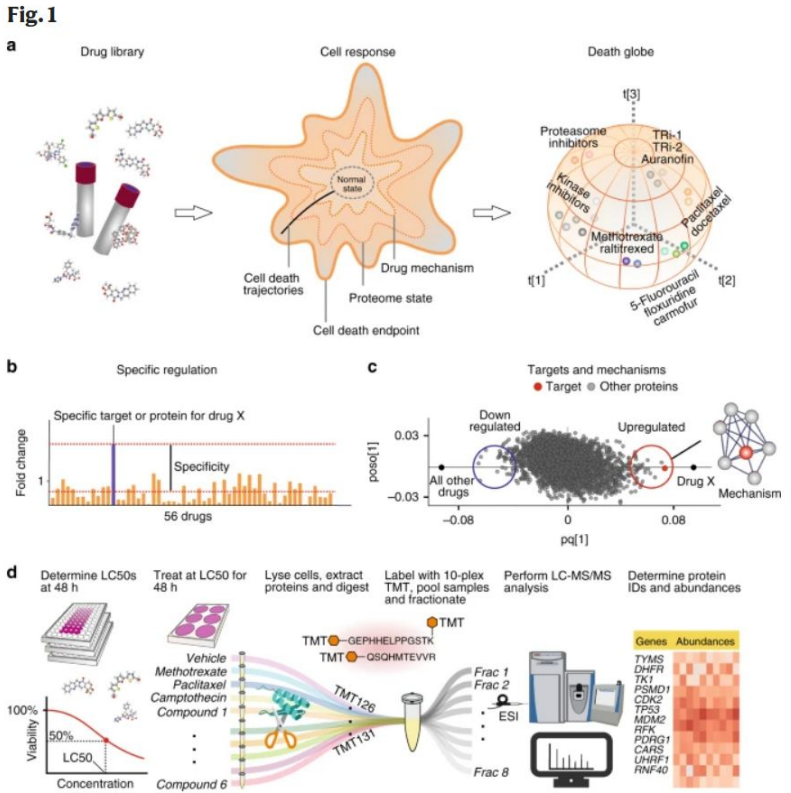
ProTargetMiner as a proteome signature

Previous study results: **Proteome signatures cluster by compound target and action mechanism of alive cells** using OPLS-DA (supervised modelling), PCA and hierarchical clustering.

New study: **Separate cell proteome signature based on treatment** (cancer medication + control) **and state** (surviving and dying cells).

Initial partial goals:

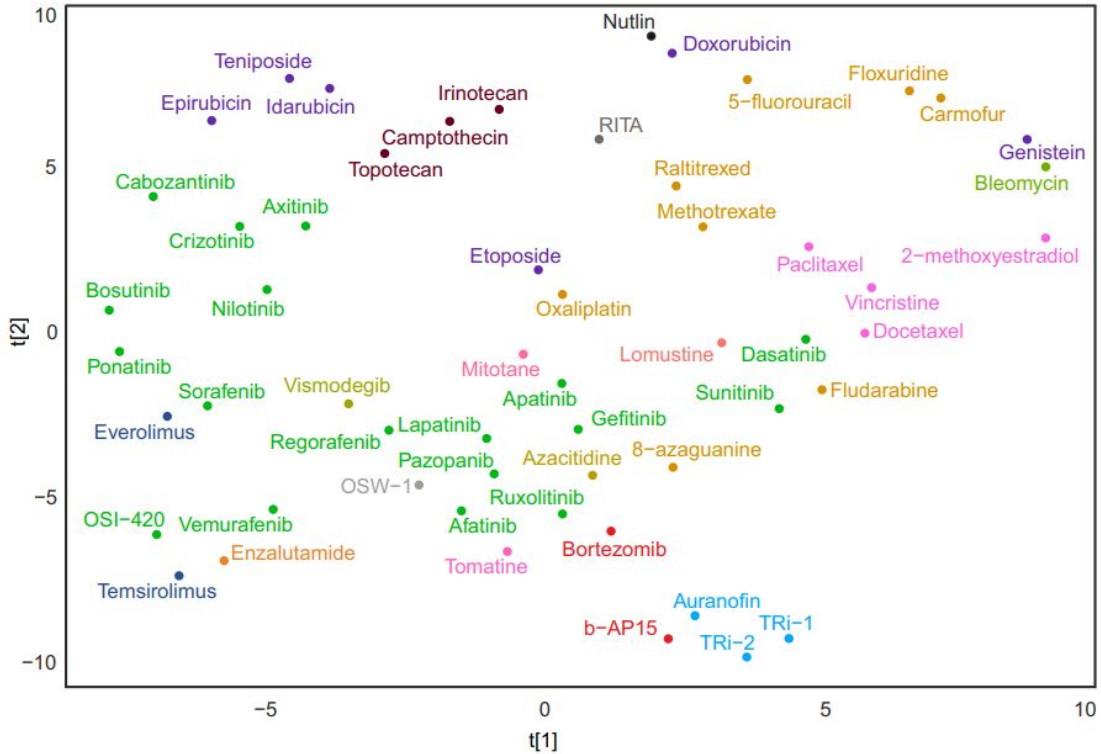
- Identifying proteome responsible for apoptosis.
- Identifying different modes of apoptosis.
- And more exciting findings... (?)



T-SNE visualisation

Drugs with similar MOA (mechanism of action) clustered. (colors)

Data is log2-transformed and normalized, but… compositional data need Aitchison transformation.



Supplementary Fig. 3 t-SNE analysis of proteome signatures. Molecules with similar mechanisms induce similar proteome changes and are found proximate on the t-SNE plot (“Death map”). Compounds in the same class (n=19) are differentiated with color. Figure made from Supplementary Data 1.

Aitchison transformation

Centered Log-Ratio (Aitchison transformation) log-ratio transformations are capable of removing the unit-sum constraint of compositional data, allowing ratios to be analyzed in the Euclidean space.

$$CLR(w^{(j)}) = \left[\log \frac{w_1^{(j)}}{g(w^{(j)})}, \ldots, \log \frac{w_p^{(j)}}{g(w^{(j)})}\right] \in \mathbb{R}^{n \times p},$$

A p-dimensional row vector

$$w^{(j)} = [w_1^{(j)}, w_2^{(j)}, \ldots, w_p^{(j)}], \text{ where } j = 1, \ldots, n \text{ is the sample index,}$$

And g(x) is the geometric mean.

Normalization method produce different correlation structures.

Badri et al. 2018 **shows that data analysis methods that rely on correlation** (such as clustering and network inference), differ depending on the normalization schemes for microbiome analysis. They further **show that log-ratio and variance stabilization transformation provide the most consistent estimates**.

NOTE: CLR and VST produce similar distributions of correlation values (that appear closer to normal distributions) than other methods

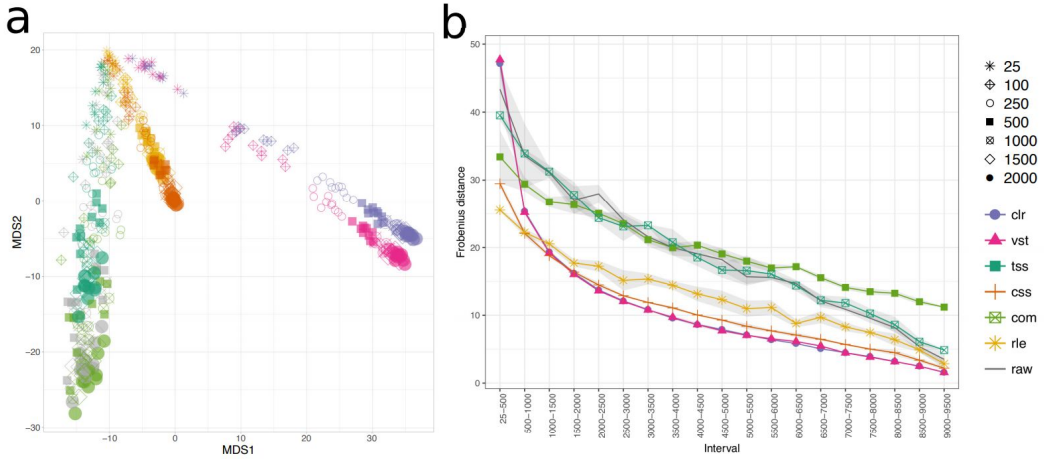


Figure 2: Transformations that remove compositional artifacts (CLR) and stabilize the variance (VST) result in substantially different patterns of correlation. A) Multidimensional scaling representation of Frobenius distance between correlation structures of varying size estimated from different normalization methods. These estimates are also compared to untransformed or raw count values (dark grey points). B) Frobenius distance between sub-samples of different sizes. Lines represent mean and grey ribbon represent standard deviation from the mean. (color scheme as in A)

Network of CLR vs TSS.

$$TSS(w^{(j)}) = \left[\frac{w_1^{(j)}}{m^{(j)}}, \frac{w_2^{(j)}}{m^{(j)}}, ..., \frac{w_p^{(j)}}{m^{(j)}}\right] \in \mathbb{S}^{n \times p},$$

The total OTU (operational taxonomic unit) count for sample j:

$$m^{(j)} = \sum_{i=1}^{p} w_i^{(j)}$$

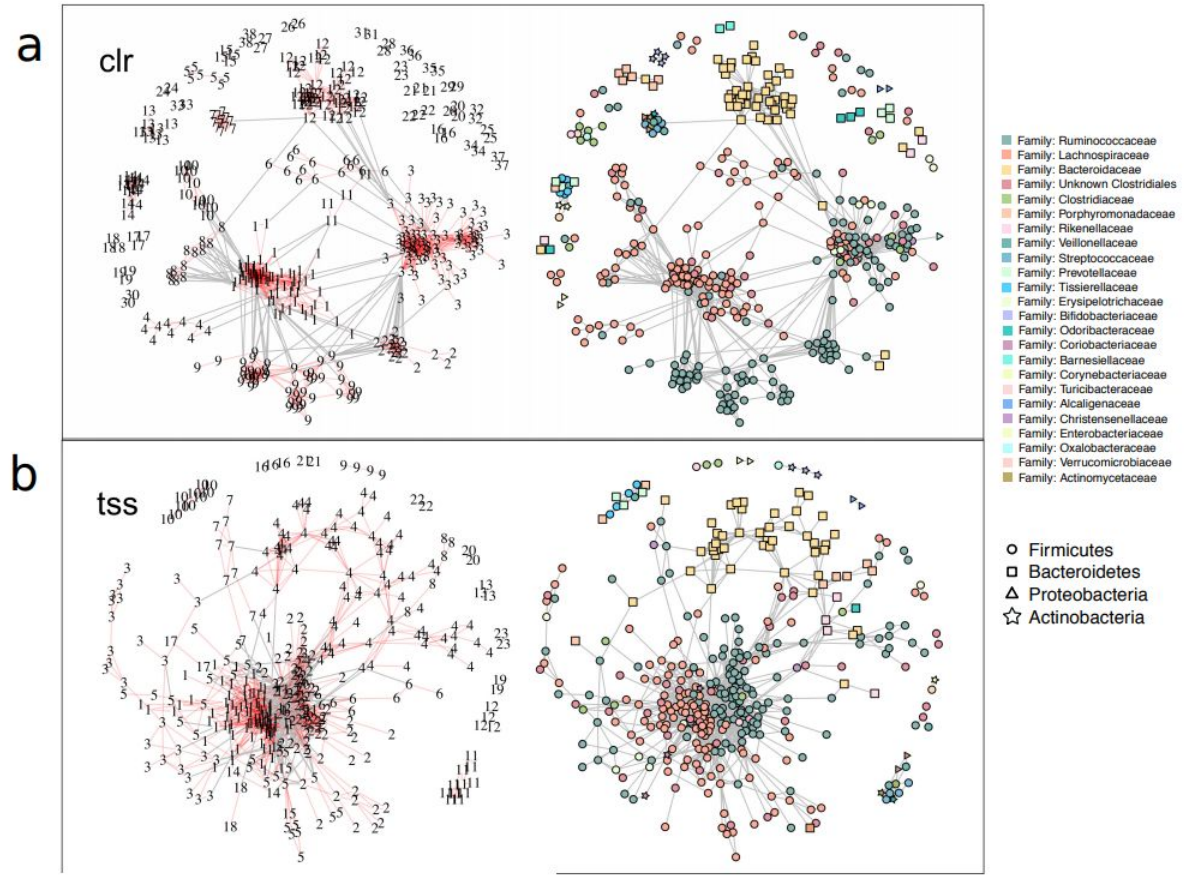


Figure 6: CLR produces more modular networks that separate communities at the Family level than TSS. For networks on the left of each panel every node represents an OTU labeled with module annotation as predicted by the Fast-Greedy modularity algorithm. The networks on the right represent the corresponding taxonomic annotation of the OTU at the family level. Layout using the force-directed Fruchterman-Reingold algorithm was conserved for both networks in each panel for comparison

More results.

Assortivity constant has something to do with phylogenetic assortativity.

Point being CLR - VST is more stable than other methods for compositional data.

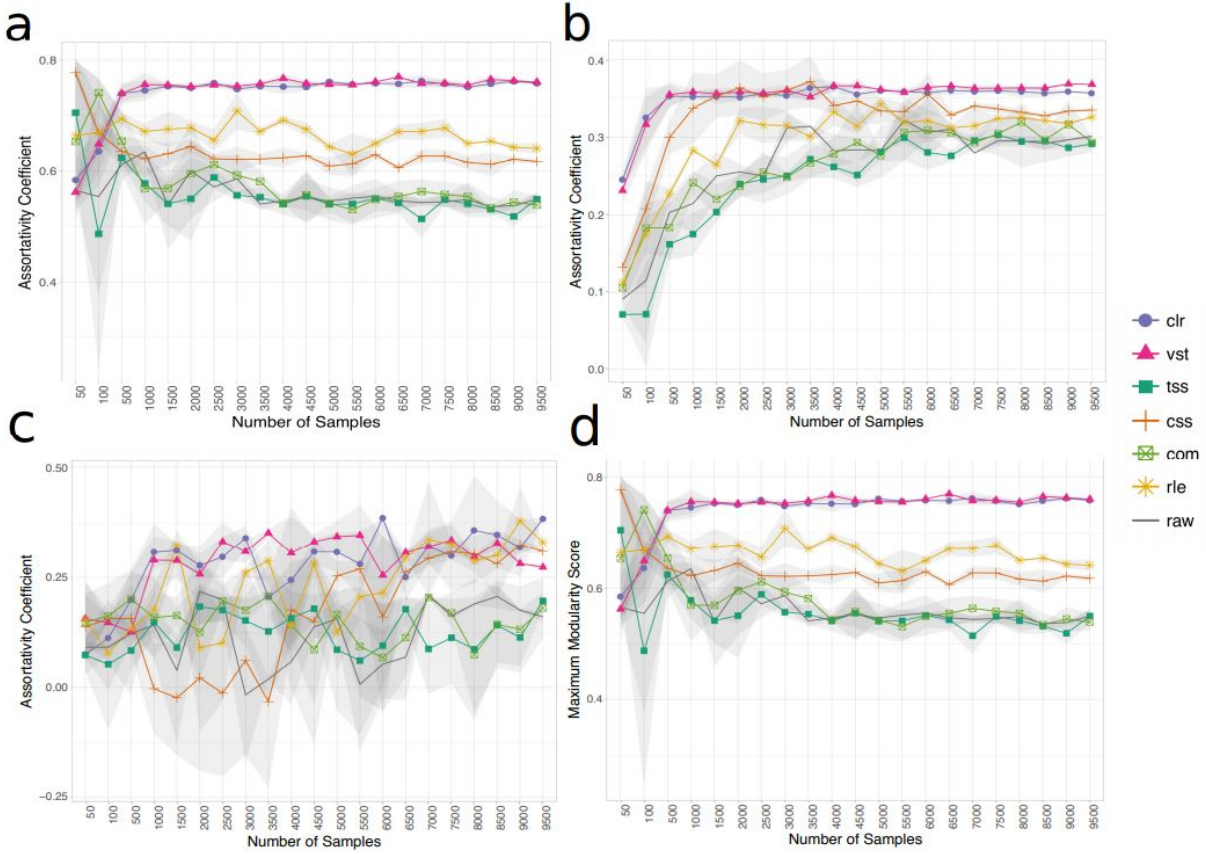


Figure 7: Relevance networks produced by CLR and VST transform contain more inter-family links independent of sample size at the top 2000 edges. A) Assortativity coefficient across sample size of family annotation. B) Assortativity coefficient across sample size of family annotation of only edges within modules. C) Assortativity coefficient across sample size of family annotation of only edges between modules. D) Maximum modularity score across sample size at 2000 edges. For all plots lines represent mean and grey ribbon represent standard deviation from the mean.

Conclusion

Results from ProTargetMiner could potentially change with transformation.

Normalize compositional data with CLR or VTS.

And a question:

- Why does it differ to use CLR-transformation than just log2-transformation?

More on practical composition statistics.

Scikit-bio CLR transformation for omics data.

http://scikit-bio.org/docs/0.4.2/generated/skbio.stats.composition.html

Medium article exploring compositional data.

https://towardsdatascience.com/exploratory-analysis-of-compositional-data-case-study-from-agricultural-soils-a1fc5f076ddc