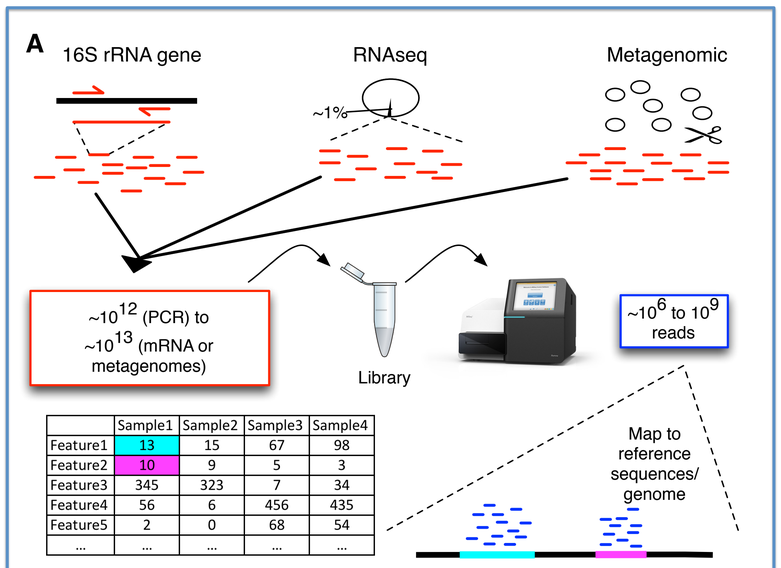
It's all relative: compositional data analysis of microbiome datasets

Greg Gloor
Deparment of Biochemistry, The University of Western Ontario
ggloor@uwo.ca

Abstract

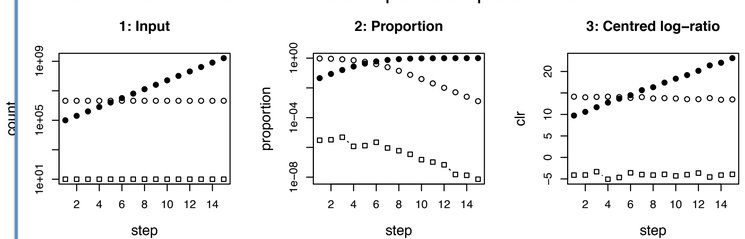
Data collected using high throughput sequencing (HTS) methods are sequence reads mapped to genomic intervals, and are commonly analyzed as either 'normalized count data' or 'relative abundance data'. One reason for these normalizations is to attempt to compensate for the problem that the sequencing instrument imposes an upper bound on the number of sequence reads. Positive data with an arbitrary bound are 'compositional data' and are subject to the problem of spurious correlation (1,5,7). Thus, ordination, clustering and network analysis become unreliable. A second problem is that the data are sparse: i.e., contain many 0 values. A third problem is that the largest measurement error is at the low count margins in these datasets (2). We use the HMP oral microbiome dataset to show how Bayesian estimation combined with compositional data approaches that examine the ratios between taxa give robust insights into the structure of microbial communities.



	Sample1	Sample2	Sample3	Sample4
Feature1	13	15	67	98
Feature2	10	9	5	3
Feature3	345	323	7	34
Feature4	56	6	456	435
Feature5	2	0	68	54
...	...	...	...	...

High Throughput Sequencing Distorts the Data

- HTS data are inherently compositional because the capacity of the machine determines the number of reads that are returned, and this is always lower than the number of molecules sampled.
- Compositional data have very peculiar properties and require special care to analyze. We would get the wrong inference about changes or correlations in the data if we compared step 1 and 15

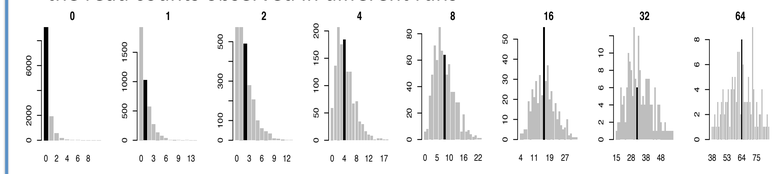


- Panel 1: Example input for sequencing where 100 features (white dots) are held constant and 1 feature (black dots) increases 2-fold for each step.
  This is the data we want to examine
- Panel 2: Shows the data after transforming sequenced reads into reads per instrument output
  - These data should not be evaluated by standard methods (1)
  This is the data that most approaches use for analysis
- Panel 3: The original shape of the data can be (often) reconstituted using the centered log-ratio transformation
  - However, the data are now ratios between the actual count and the geometric mean count (gx), so interpretation must be cautious

$x=[x_1,x_2 \dots x_n]$ $\quad clr_x=[\log(x_1/gx), \log(x_2/gx) \dots \log(x_n/gx)]$
This is the data we should use!

Better to think of probabilities not counts

- Random sampling of the environment and the libraries causes variation in the read counts observed in different runs



- For a given read count (labeled top of histogram) the black bar represents the abundance of the count in replicate 1, the grey bars indicate the distribution of counts for the same features in replicate 2
- Replicates are aliquots of the same library run on different lanes
- These data are approximately multivariate Poisson distributed, this can be estimated in a Bayesian framework using Dirichlet sampling prior to clr transformation. (2,3)

Analysis Goals

Given a set of data derived from HTS:

1) Is there structure?
2) Are there differences between groups?
3) What correlates?

Practical approaches to the 0 problem

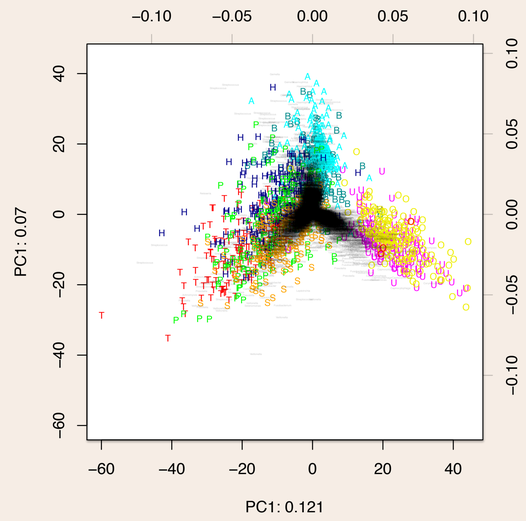
- Values of 0 are not compatible with compositional data analysis.
- When an OTU has 0 in all samples, it is likely that the value of 0 indicates that the OTU **cannot** be observed. OTUs that are 0 in all samples should be removed
- When an OTU has a value of 0 in some samples and a value greater than 0 in others, the value of 0 most likely means that the OTU **could have been** observed. In this situation we must estimate the background frequency of the count.
- All observed counts are probabilities of observing the count given their true frequency in the sample and the sequencing depth.
- Compositional biplots can be generated using point-estimates where the count zero multiplicative correction (6) or a pseudocount of 0.5 is applied to 0 values.
- Between-group differences and the ϕ-statistic can be calculated as expected values of the statistic computed from the distribution of probabilities compatible with the observed count estimated by sampling from a Dirichlet distribution prior to clr transformation.

Conclusions

- Compositional data analysis methods are fully compatible with microbiome datasets generated by HTS.
- Values of 0 are problematic, but can be addressed.
- β-diversity can be easily measured using Singular Value Decompositions and displayed using a biplot rather than distance-based metrics.
- Univariate differences and correlation can be examined as expected values of effect sizes and ϕ, a measure of the concordance of ratios.

1) Aitchison,(1986) The statisitical analysis of compositional data (Blackburn Press) and Pawlowsky-Glahn (2015) modeling and analysis of compositional data (Wiley)
2) Fernandes (2013) ANOVA-like differential expression analysis for mixed population RNA-seq data (PLoS ONE)
3) Fernandes (2014) Unifying the analysis of high throughput sequencing datasets (Microbiome)
4) Gloor (2016) Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes (J. Comp. Graph. Statistics)
5) Lovell (2015) Proportionality: A Valid Alternative to Correlation for Relative Data (PLoS Comp. Bio)
6) Fernandez (2014) Bayesian-multiplicative treatment of count zeros in compositional data sets (Statistical Modeling)
7) Friedman (2012) Inferring correlation networks from genomic survey data (Plos Comp. Bio)

Structure: β-diversity via the compositional biplot

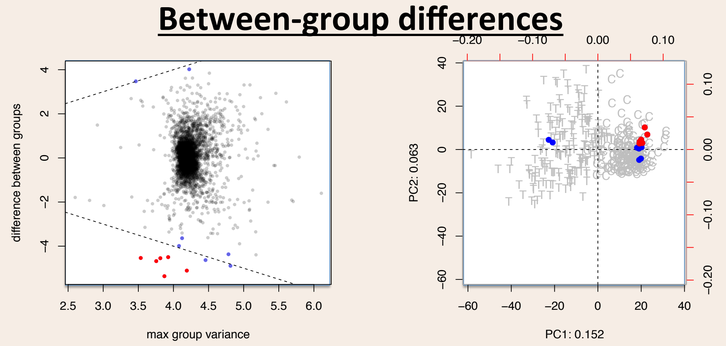


Compositional biplot showing the variance structure of the entire oral microbiome dataset. The gross structure of the oral microbiome splits into three main groups (from the right and proceeding counter-clockwise): the sub and supra gingival plaque (U and O); most samples from attached gingiva (A) and buccal mucosa (B); most samples from tongue dorsum (T), tonsils (P), saliva (S). The hard palate (H) appears to be split between these latter two groups. It is noteworthy that these groups largely correspond to the location of the sampling sites in the oral cavity. Samples are colored according to the sampling site with the OTU variance overlaid on top. The dataset has 1446 samples and 5203 OTUs. Explaining 19.1% of the variance in the first two components suggests the data are relatively robust for such a large dataset. However, it would not be appropriate to use these results to make strong conclusions about relationships between OTUs in the data because there could be latent associations. Note that this biplot is drawn to display, as much as possible, the actual relationships between the samples (scale=0).

Site	Grp 1	Grp 2	Grp 3
O	4	0	182
U	5	1	187
A	184	0	31
B	166	7	6
H	96	78	1
T	36	151	0
P	71	109	4
S	56	106	3

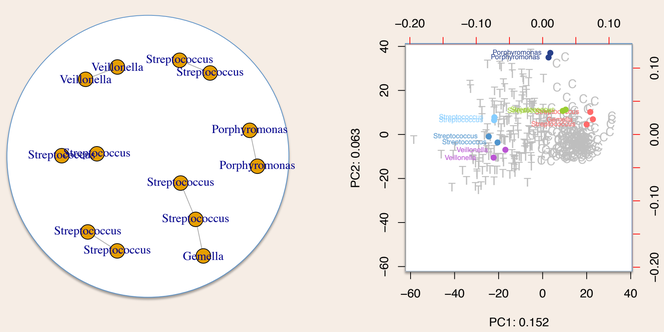
K-means group membership. Not surprisingly, the plaque samples (O, U) are most separate from the rest, since they are essentially sequestered away from the others. Next most separate are the buccal mucosa (cheek) and attached gingiva (B,A), these are in almost constant contact. The remainder of the sites are much more variable in their group membership, but largely fall into a separate group.

Between-group differences



The left panel shows an effect plot (4) and the right panel shows a compositional biplot of the tongue dorsum (T) and buccal mucosa (C) subset illustrating between-group differences. Each point in the effect plot shows the denominator and numerator of an effect size statistic, which is a more reliable indicator of difference than is a p-value. The red points show OTUs with effect sizes greater than 1, and blue points show OTUs with an effect greater than 0.8. The vast majority of effect sizes are very small, indicating trivial OTU abundance differences between the two groups. Values were calculated from the posterior distribution of OTU probabilities generated from a Dirichlet process, with the distributions being transformed by the clr prior to analysis using the ALDEx2 Bioconductor R package. The biplot shows that the OTUs with the largest effect are among the most variable in the dataset.

Correlation



The left panel shows clusters of OTUs with a symmetric ϕ-statistic (5) of 0.2 or lower, and the right panel shows a compositional biplot of the tongue dorsum (T) and buccal mucosa (C) subset illustrating between-group differences. The symmetric ϕ-statistic is a measure of the constancy of variance between OTUs across all samples. With one exception in this dataset, OTUs with low ϕ values are found to be classified with the same taxonomic name. This likely indicates incomplete clustering of OTUs. Values were calculated from the posterior distribution of OTU probabilities generated from a Dirichlet process, with the distributions being transformed by the clr prior to analysis using the ALDEx2 Bioconductor R package. The biplot shows that the OTUs with the largest effect are among the most variable in the dataset.