

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

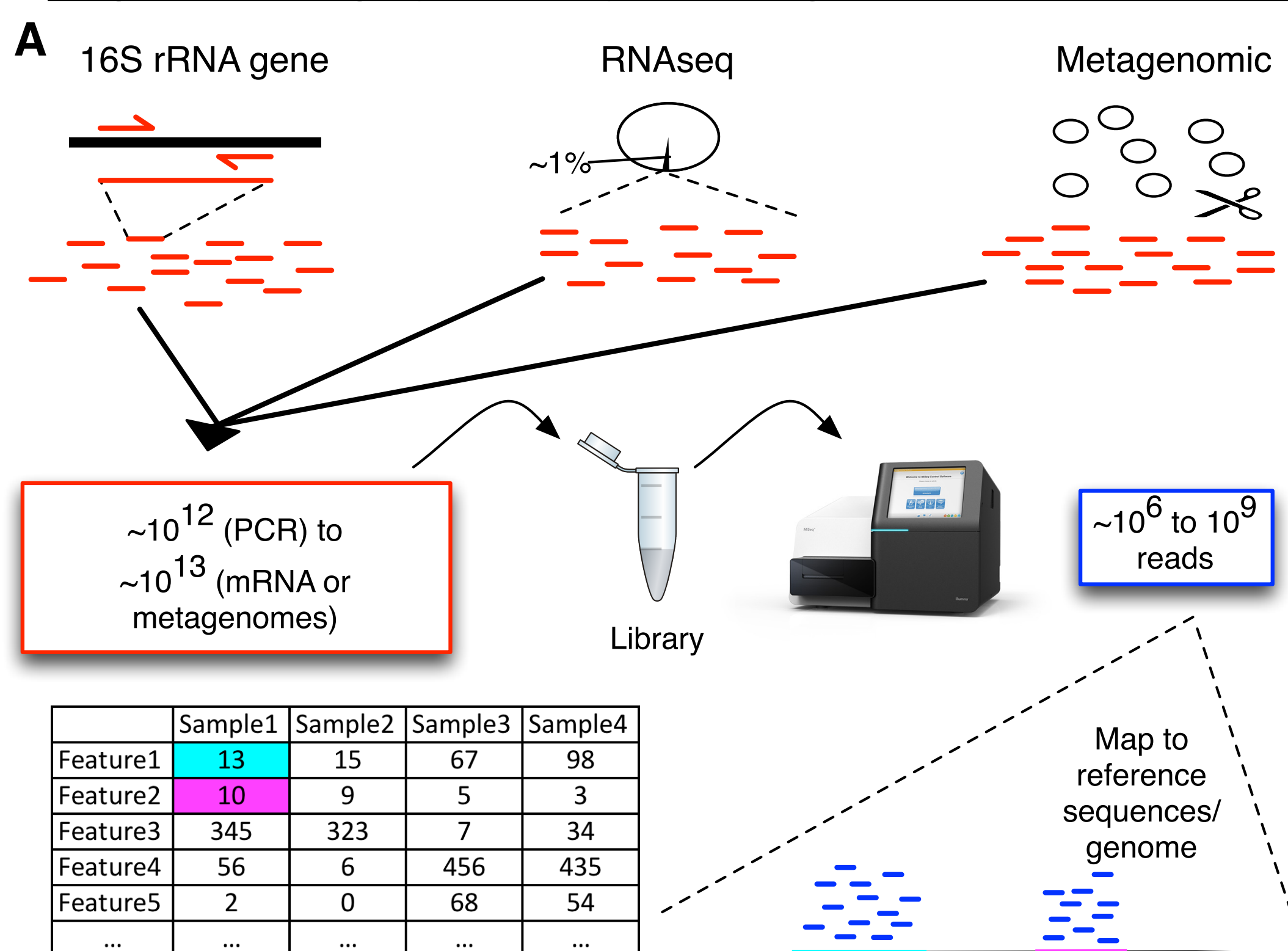
Greg Gloor

Department of Biochemistry, The University of Western Ontario  
[ggloor@uwo.ca](mailto:ggloor@uwo.ca) • [ggloor.github.io](https://github.com/ggloor) • [github.com/ggloor/CoDaSeq](https://github.com/ggloor/CoDaSeq)

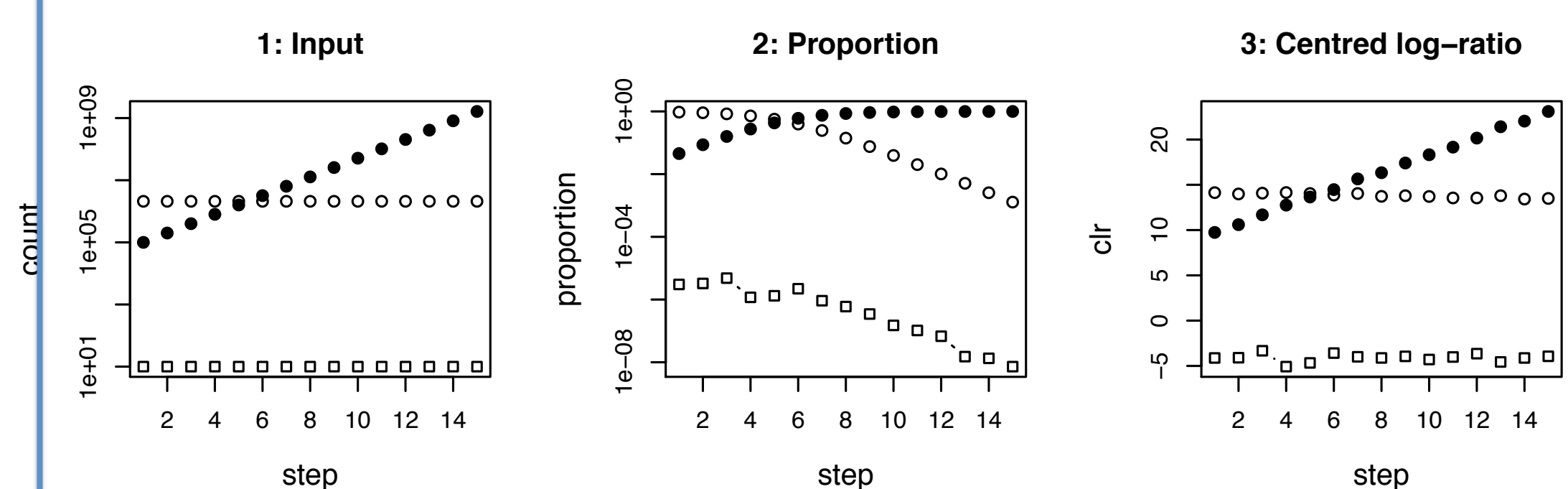
## Abstract

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' with or without normalizations for sequencing depth. This is done in an 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' (CoDa), where the datapoints are non-independent and so 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). These issues result in unstable and irreproducible analyses. We use a subset of 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.

## High Throughput Sequencing Distorts the Data



- HTS data are inherently compositional because machine capacity is always an arbitrary number of reads with an upper bound.
- Compositional data have very peculiar properties and require special care to analyze (1). We would get the wrong inference about either change or correlation in the example data below if we compared samples collected from step 1 and 15 from the 3 panels.



- 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 think we are examining
- Proportions shows transformed sequenced reads from panel 1 as 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

$$x = [x_1, x_2, \dots, x_n] \quad \text{clr}_x = [\log(x_1/g_x), \log(x_2/g_x), \dots, \log(x_n/g_x)]$$

This is the data we should evaluate!

## Common Analysis Goals

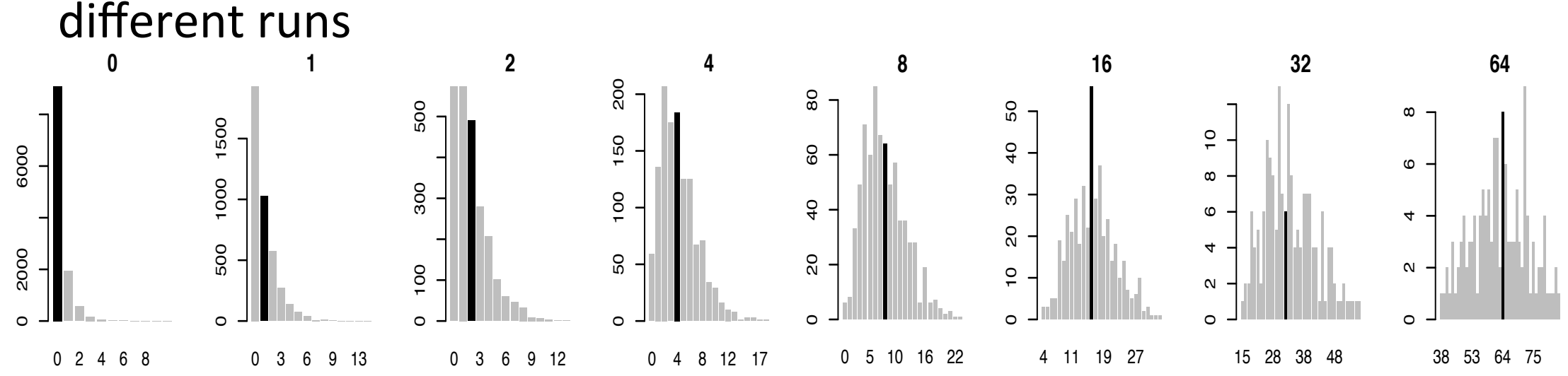
Given a set of data derived from HTS:

- 1) Is there structure?
- 2) What are the differences between groups?
- 3) What correlates?

## Are values of 0 compatible with CoDa analysis?

- When an OTU has 0 in all samples, it is likely that the OTU **cannot** be observed and should be removed.
- When an OTU has a value of 0 in some samples and a value greater than 0 in others, the OTU **could have been** observed. Here, we must estimate the background frequency.
- All observed counts can be thought of as 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 is applied to 0 values.
- Between-group differences and the  $\phi$ -statistic can be calculated as expected values of the statistic computed from the clr transformed posterior distribution of probabilities estimated by sampling from a Dirichlet distribution.

## Better to think of probabilities not counts

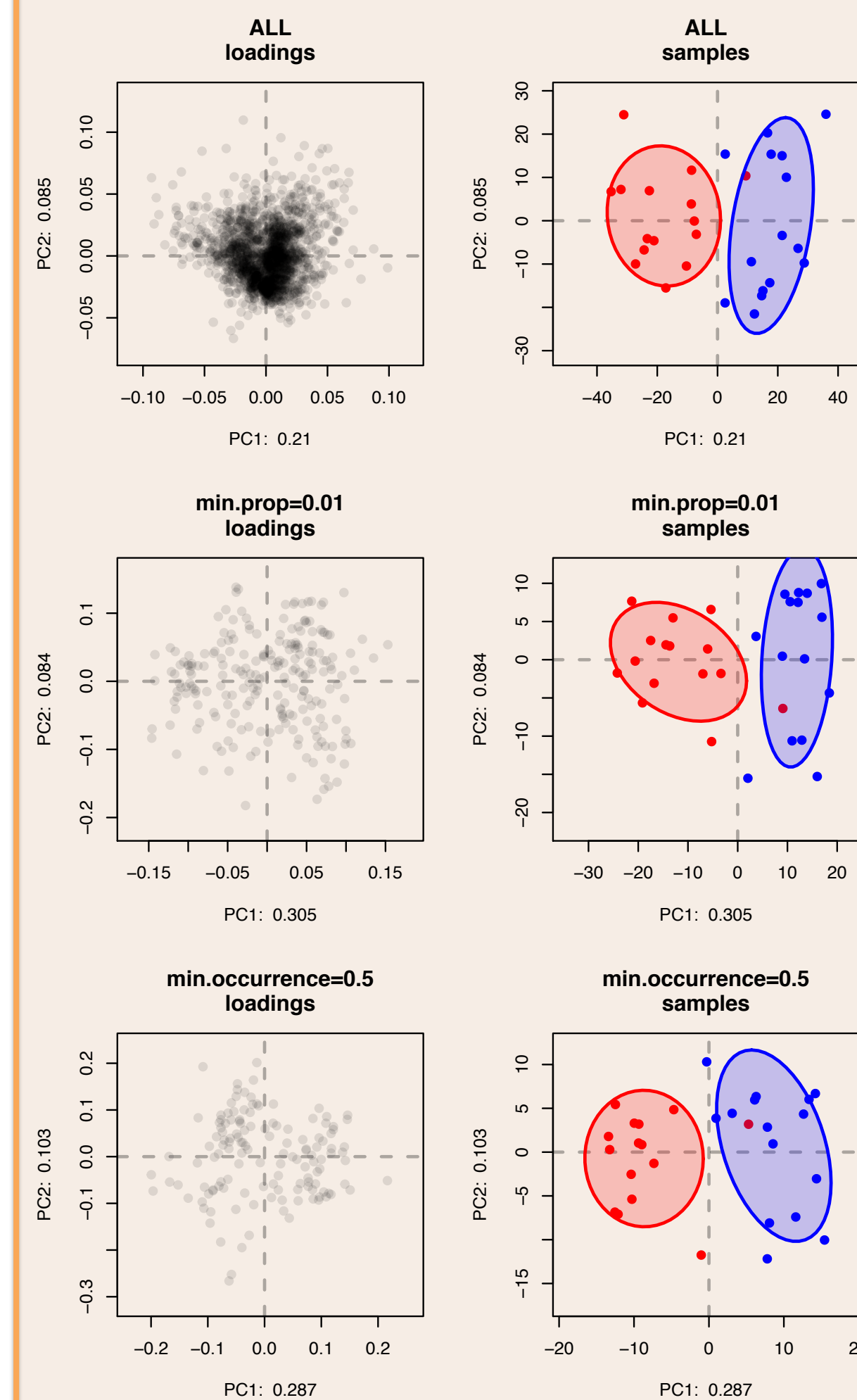
- Random sampling causes variation in the read counts observed in different runs
- 
- For a given read count (labeled top of histogram) the black bar represents the abundance of the count in replicate 1, the grey bars indicate the distribution of counts for the same features in replicate 2
  - Replicates 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)

## Conclusions

- Compositional data analysis methods are fully compatible with microbiome datasets.
- OTUs with 0 can be addressed (6,8).
- $\beta$ -diversity can be easily measured after Singular Value Decomposition and displayed using a variance-based PCA plot of rather than distance-based PCoA plot.
- Univariate differences and correlation can be examined as expected values of effect sizes and  $\phi$ , a measure of the concordance of ratios.
- Remember that variance and not abundance is being examined.
- Using compositional approaches for the three analysis objectives returns consistent data that is minimally affected by filtering or other changes in the analytical protocol.
- All tools return outputs that can be readily interpreted in the context of any of the other tools. For example, the OTUs that are most different between groups are seen to be the ones that have the largest loadings in the PCA plots, and the association test identifies groups of taxa that are changing reproducibly in the same direction.

- 1) Aitchison, (1986) The statistical 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)
- 8) Gloor (2016) Compositional uncertainty should not be ignored in high-throughput sequencing data analysis (Aus. J. Stat)

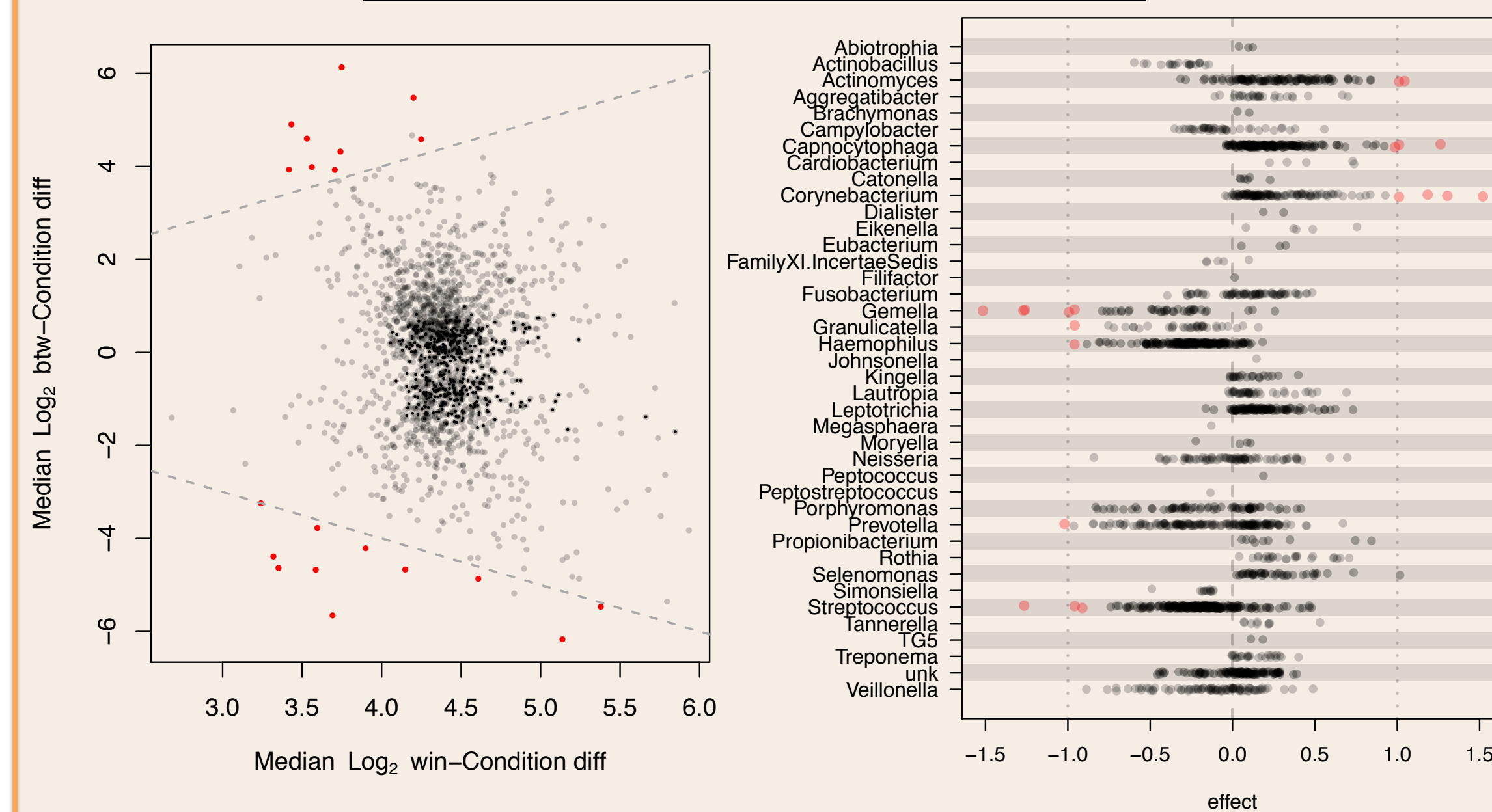
## Consistent $\beta$ -diversity via variance not abundance



Compositional PCA plots showing the variance structure of 15 AK and OP samples of the HMP dataset samples chosen at random. The loadings plots on the left show the contributions of each OTU to the separation of the samples on the right. The location of the OTUs relative to the origin is a linear function of their standard deviation in the dataset.

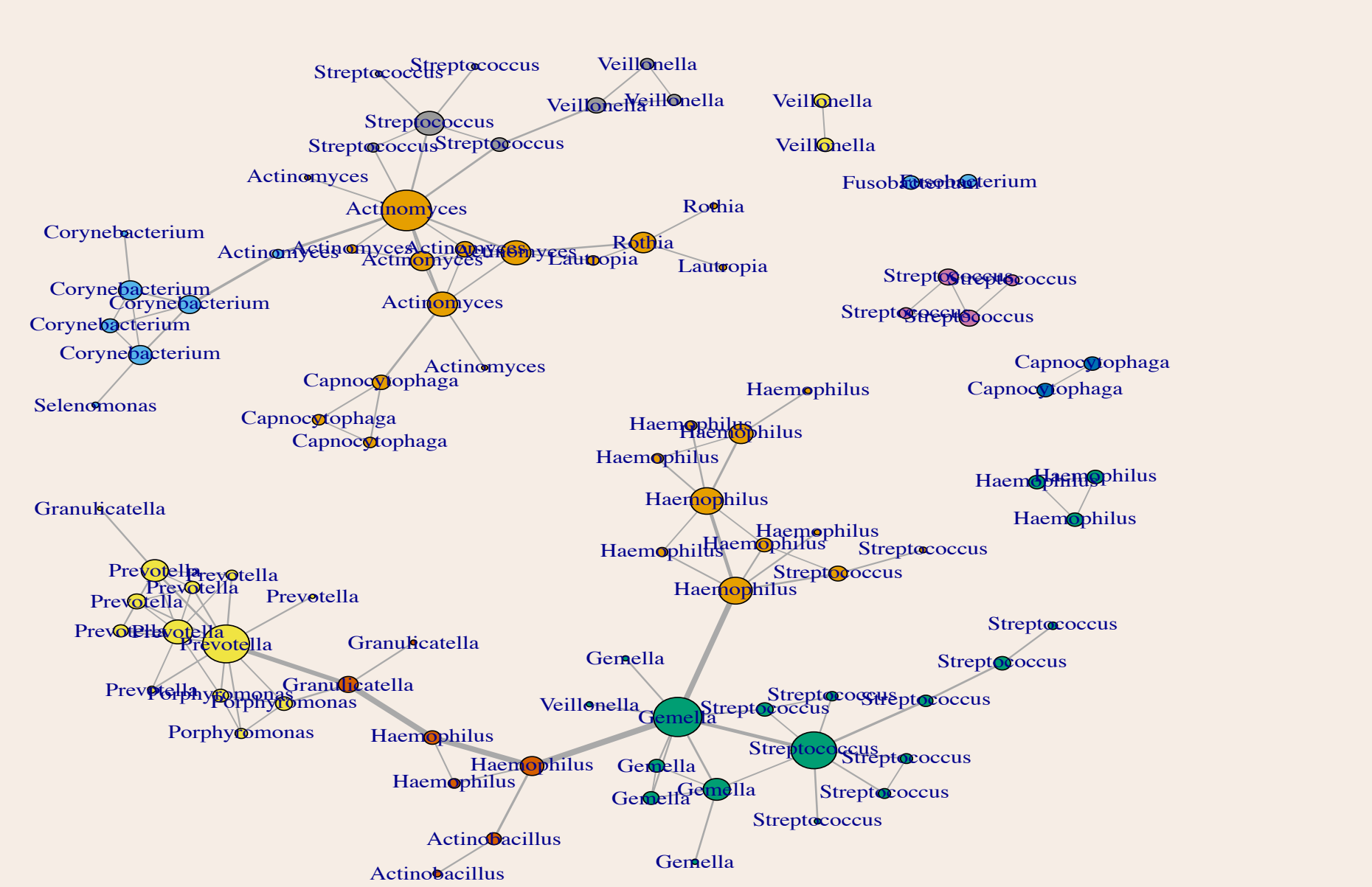
Note that the ordination is robust to filtering the. The separation of the samples is **not** based on abundance of an individual OTU, but is based on the variance in the ratios of abundance between all OTUs in the subset.

## Between-group differences



The left panel shows an effect size plot (4) and the right panel shows a plot of the OTUs grouped by genus with the effect size on the x axis. The red points show OTUs with an expected FDR < 0.05, otherwise OTUs are colored in grey. The vast majority of effect sizes are very small, indicating trivial differences between the two groups, because the variation within groups is very large. Note how most OTUs in a genus tend to change in the same direction. Positive values indicate OTUs that are relatively more abundant in AK than in OP. Values were calculated using the ALDEx2 Bioconductor R package (2,3). OTUs were filtered to include only those with a frequency > 0.001 in any sample.

## Compositional Association



Determining correlations within compositional data is fraught with problems (1,5,7). The only method independent of assumptions is to identify OTUs with nearly constant ratios across samples. This can be done with the  $\phi$ -statistic (5), and the graph above shows the connectedness of taxa with a  $\phi$ -statistic less than 0.3.