And I forgot to mention but my personal preference would be to do Bray Curtis and DPCoA.

Ellie Margolis

**From:** "Margolis, Ellie" <[Ellie.Margolis@STJUDE.ORG](mailto:Ellie.Margolis@STJUDE.ORG)>  
**Date:** Wednesday, October 16, 2019 at 2:43 PM  
**To:** "Tang, Li" <[Li.Tang@STJUDE.ORG](mailto:Li.Tang@STJUDE.ORG)>, "Sun, Hongying" <[Hongying.Sun@STJUDE.ORG](mailto:Hongying.Sun@STJUDE.ORG)>  
**Cc:** "Alsallaq, Ramzi" <[Ramzi.Alsallaq@STJUDE.ORG](mailto:Ramzi.Alsallaq@STJUDE.ORG)>  
**Subject:** Background on Beta Diversity

Hi Li and Hongying-

I promised I’d send some reading material about different methods to compare two microbiome communities (beta diversity).

If the question is how dissimilar is this community from another (in our case the mock ‘true’ from the mock ‘classified with a given ref set’) there are four approaches I’m aware of:

1. Non-phylogenetic/Non-taxonomic Counting-> weighted Bray Curtis
2. Phlyogenetic or taxonomic parsimony-  ask how many internal nodes of community tree are shared -> weighted unifrac
3. Generalized principal component analysis- tree structure incorporated to find components of species abundance -> DPCoA
4. Phlyogenetic probability distributions- takes into account uncertainty in read location and then compares probability distribution over tree -> phylogenetic Kantorovich-Rubinstein

The disadvantage with 1 is that it often contains errors introduced b the naming process and losses the detailed information that phylogeny offers.   The disadvantage with 2 is that it is extremely sensitive to noise. In principal both 3 and 4 are the most sound but most computationally intensive.  I’ve attached articles about the latter two.

Ellie Margolis MD PhD

Assistant Member

Infectious Diseases Department

St Jude Children’s Research Hospital