**Albanese et al. Explaining diversity in metagenomic datasets by phylogenetic based feature weighting.**

i) ANOSIM is used as an analysis of similarity. As I understand it, it compares the variation of beta diversity between classes - not the beta diversity differences themselves. PERMANOVA is used as a more empirical measure of the differences in beta diversity between subclasses. With that in mind, it is quite suspectful that in general, when assessing the impact of different number of clades, there is persistently a lower correlation (0.13-0.44) from the PERMANOVA test than there is from the ANOSIM test (0.27-0.68). Furthermore, the optimal clade number selection has been indicated to be modeled on the highest correlation values for PERMANOVA and ANOSIM. Given that ANOSIM assesses correlations in the *variability* of beta diversity values between the clades, it seems erroneous to be using the high(er) correlation from ANOSIM as a means of selecting the number of clades. Using that, all you get is the optimal number of clades where members show similar changes in their beta diversity, not the optimal number of clades where members can be distinguished by their beta diversity.

ii) The PCA analysis of the weighted UniFrac distances indicated Venezuela samples’ distributions were intermediary to the “Western” and African samples. I do not understand why the authors then chose to club together Venezuela with the African samples and carry forward with the classification using that 2 class system. I wonder if it would’ve been worthwhile to treat Venezuelan samples as a third set and see where they get placed when the clades have been weighted and the subtrees ranked according to the PhyloRelief algorithm.

iii) The authors point out the differentiation between the samples based on age. They further emphasize how age has a greater influence in the Western samples as compared to the non-Western samples. Can this also not be due to the fact that they have 316 western samples (and 225 non-western ones), and consequently a larger set of OTUs from the Western samples (that can then be used as a discriminant during clade separation).

iv) The number of clades chosen by the PERMANOVA/ANOSIM tests to differentiate individuals above and below age 2 (for Western and non-Western samples) was 90 and 30 respectively. The input data had 98 families (316 samples) from USA and Italy (western), and (34+19) families (225 samples) from non-western geographical locations. I find it interesting that the number of families land up so close to the optimal number of clades established by this approach. It appears that the discrimination by the metagenome is overwhelmed by the discrimination derived from the sample families. I am curious if this is an instance of batch effect showing through, and if that is not the case, what would be a method by which we could assess the same.

**Imelfort et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes**

i) When binning the filtered contigs using the TF-ESOM approach, it seems quite odd that the classification groups were established by manual identification of boundaries within the U-matrix. Couldn’t a rank based decomposition of the U-matrix be used as one mathematically robust approach for classification instead? Manual identification sounds extremely handwavy, especially when they are using this output as a comparison to GroopM’s performance. The authors point out that GroopM’s errors were localized to seven genome bins, whereas TF-ESOM errors appeared ‘more systemic’ (61 bins) - perhaps the manual bin assignment might be the ‘systemic’ cause they are thinking of.

ii) The authors use a similar hand-wavy ‘assessment of the graphics’ approach to carry out merging and deletion actions on the groups generated by GroopM. The process appears to be iterative and manual, which does not speak very well to automated implementation of the package by other users. They subsequently use the same approach when doing the performance comparison with SPAdes on the Sharon dataset. How replicable and scalable is this sort of implementation?

iii) In the modification of the coverage profiles to remove well-studied strains, the authors make no note of normalize the samples after the removal of such strains. I am wondering, since the method looks at variance in coverage profiles for the different OTUs, whether a one-stroke removal of IMG species without compensating for the change in overall coverage in the respective samples (although there are just 3 samples) could have any impact on the efficiency of binning by GroopM?

**Suzuki et al. Faster sequence homology searches by clustering subsequences.**

i) Why does the length of seed subsequences need to be restricted to 6-8 residues for a perfect hash function? A perfect hash function can accommodate unique representation of ‘n’ residues, corresponding to n distinct integers. Since we are using a perfect hash, the chances of randomly matched cases (or collisions) are 0 anyways, and hence using longer subsequences for seeds should not have a negative impact on search sensitivity. Could there a more rational way of assessing which subsequence length to restrict oneself to?

ii) Isn’t it a bit harsh to use the reduced alphabet alongwith the restriction of 6-8 residues for the seed subsequence in this method? Furthermore, using the max(BLOSUM62 score) for a cluster of AA represented in the reduced alphabet does not seem suitably justified. Why not take the mean as a measure? It would appear that with the max() approach, we impart equal biological significance to proteins sharing 62% similarity in their side chain structure - which may not necessarily be true, nor a biologically accurate way of going about clustering different AA sequences.