At the outset of doing the multivariate analysis on high-throughput sequencing raw data from DNA, RNA, and microbiome, which observed that contain many zero values. These zero may cause false discovery rate in analyses in difference gene expression. In order to derive meaningful results from raw data, the most challenging part is looking for best model to analysis the characteristic zero values.

Before designing models for handling zero values, should be knowing the concept of Zero Generating Processing (ZGP). ZGP has three major classing: Sampling Zero, Biological Zero, and Technical Zero. The Technical Zero has sub size, partially and completely.

After Silverman used simulation to studies given models, they conclude their designed models and ranked models for best performance. The top ranked is Random Intercept (RI) model, it is best performed on sampling zero and it is designed used to model sampling and partial technical zeros at initial. Second best performing models are Base and Biological Zeros (BZ) models; the base model is following the Poisson distribution and Biological Zero is designed on modeling biological zero arise due true biological absence. Base and BZ are seeking to use for sampling zero and Base similar results as RI, but BZ has well performance on biological zero. Third, Pseudo-Count (PC) and Zero-Inflated Poisson (ZIP) models are poorly performances.

Zhou introduced a transformed multivariate modeling for microbial counts, it’s zero-inflated Dirichlet-tree multinomial (ZIDTM) model. From raw microbiome data, utilities Operational Taxonomic Units (OTU) to build phylogenetic tree, which comes with one root, and it’s developed more and more nodes following a root, each node is independent. In additional, ZIDTM is product of zero-inflated beta-binomial (ZIBB) model, ZIBB is reduced from zero-inflated Dirichlet-multinomial (ZIDM) model, and ZIDM is based on Dirichlet-multinomial (DM) model. As use ZIDTM model in analysis zero, then it must follow test result by specific method based on DA testing. After Zhou’s group done the simulation, they conclude adaANCOM is the best methods in DTM and ZIDTM models and ZIDTM is most flexible multivariate distribution for count data.

The group of Tang has published, DM model cannot flexibly handle various dispersion patterns even though the zero-inflation. Then they develop it to generalized DM (GDM) which allowing more general covariance structure. In addition, to develop new probability distribution for supply the over-dispersion and zero-inflation of data, zero-inflated GM (ZIGDM). After the simulation of gut microbiome data, they have demonstrated, GDM is superior fit to taxon count to DM and ZIGM is improve the goodness fit for taxa with zero counts. As facing the number of taxa is larger than sample size, the multivariate association test cannot handle high-dimensional microbial taxa. However, the ZIGDM can fit in incorporate standard regularization approaches to deal with high dimensionality.

In Jiang’s paper, introduced a statistical method for taxon count while microbiome data is non-biological zero, the method named mbImpute. The major idea of mbImpute is borrow information and to impute non-biological zeros in microbiome data. Before the imputation, mbImpute requires taxon’s counts on the same scale. However, the taxon’s counts are not on same scale then users may face challenging on normalize microbiome data and mbImpute has a default normalization which normalizes samples by library size. The mbImpute is doing two step for done the imputation: first is identification taxon abundance and once identifies abundances are likely missing using a mixture-modelling approach, then second is imputation of missing taxon abundance.