Finding the centre: corrections for asymmetry in high-throughput sequencing datasets

Jia R. Wu[1], **Jean M. Macklaim**[1], **Briana L. Genge**[1], **and Gregory B. Gloor**[1,2]
[1]Dep't of Biochemistry, U. Western Ontario, London, Canada, N6A 5C1
[2]Dep't of Applied Mathematics, U. Western Ontario, London, Canada, N6A 5C1
gbgloor@gmail.com

Abstract

High throughput sequencing is a technology that allows for the generation of millions of reads of genomic data regarding a study of interest, and data from high throughput sequencing platforms are usually count compositions. Subsequent analysis of this data can yield information on transcription profiles, microbial diversity, or even cellular abundance in culture. These data have many pathologies: because of the high cost of acquisition the data are usually sparse, and often contain far fewer observations than variables. However, an under-appreciated pathology of these data are their often unbalanced nature: i.e, there is often be systematic variation between groups simply due to presence or absence of features, and this variation is important to the biological interpretation of the data. A simple example would be comparing transcriptomes of yeast cells with and without a gene knockout. This causes samples in the comparison groups to exhibit widely varying centres. Despite the compositional nature of sequencing data, most tools inappropriately model the the underlying features in sequencing data as linearly independent counts. This work extends a previously described log-ratio transformation method that allows for variable comparisons between samples in a compositional context. We demonstrate the pathology in modelled and real unbalanced experimental designs that have a unidirectional direction of change to show how this dramatically causes both false negative and false positive inference in both traditional and compositional approaches. We then introduce several measures drawn from the RNA-seq and robust CoDa analysis fields to demonstrate how the pathologies can be addressed. An extreme example is presented where only the use of a predefined basis is appropriate. The transformations are implemented as an extension to a general compositional data analysis tool known as ALDEx2 or ANOVA Like Differential Expression.

Key words: transcriptome, Bayesian estimation, count composition, sparse data, high throughput sequencing, robust estimation, qPCR