**Mother’s milk and baby formula: prediction, correlation, and classification within the joint host and microbiota transcriptome structure**

Scott L. Schwartz, Iddo Friedberg, Ivan V. Ivanov, David B. Dahl, Sharon M. Donavan, and Robert S. Chapkin

A recent study sought to compare infants who were exclusively formula fed (FF) to infants who were exclusively breast fed (BF) with respect to host mRNA phenotypes and metagenomic profiles. Host mRNA microarray measurements were obtained in a non invasive manner by isolating eukaryotic material in stool samples collected from the babies at 3 months after term birth. Bacterial content from the same stool samples was additionally extracted and sequenced using a Roche 454 sequencer, thereby providing a complete picture of the integrated gut system. Previous work with the data showed clear classification between FF and BF babies on the basis of host mRNA expression levels. Subsequently, we have seen similar classification capabilities FF and BF babies on the basis of bacterial metabiome phylogenetic profiles. To begin to fully harness the synergistic information from both eukaryotic host cell gene expression and the prokaryotic bacterial colony profiles, we examined several quantitative approaches for integrating the two data sets. First, we predicted single host gene expression for 16,853 genes using percent of Firmicutes and percent Actinobacteria (two phylum that clearly distinguished formula fed versus breast fed babies) present in the gut as linear covariates. For 394 genes, all model coefficients were significant and R-squared values were generally larger than 0.7. False discovery rate (FDR) corrected likelihood ratio test q-vaules for these gene models were less than 0.25. Next, on the basis of a preliminary principal components analysis, we examined correlations between percentages of Carbohydrates, Virulence, Cell Wall and Capsule, and RNA metabolism in the metabiome (as categorized by SEED level 1 biological processes) and 519 host immunity and defense genes (as categorized by PANTHER biological processes). Strong correlations of |R-squared| > 0.7 with FDR q-values of less than 0.2 were observed for 20 genes. Finally, in addition to these results, we also present some preliminary results involving a deeper examination of the hierarchical SEED classifications relative to FF versus BF babies.