Gut-Microbiome Analysis: Finding Important Predictors within Responding and Non-Responding Patients with respect to a Ketogenic Diet.

James Ziegler

HIVE Lab (High-performance Integrated Virtual Environment)
George Washington University, Washington DC, 20052
May 14th, 2021

1. Summary
   We are interested in knowing which of the organisms in the patients' samples are important predictors of whether or not a given patient will respond to a ketogenic diet. Because the gut-microbiome of a patient is altered during a specific diet, and a known response to the diet has already been recorded, we want to use these important predictors to determine whether or not a patient will respond to the ketogenic diet before they even begin said diet. This can be accomplished using Regularized Linear Discriminant Analysis[1] (RLDA) in HIVE. We conclude from our analysis that it is prudent to evaluate at least the top 25 predictors to identify overlap between MATLAB and RLDA as implemented in HIVE.

2. Objectives
   Using RLDA in HIVE, find contribution coefficients, cumulative contribution coefficients, P-Values and Student T-Values in order to show important predictors.

3. Methods
   a. RPKM Table
      i. In HIVE navigate to Portal > Alignment Comparator (under Classifications).
      ii. For General Parameters: in the dropdown for Alignments to Use select HIVE IDs 21025, 21025, 21023, 21022, 21021, 21031, 21030, 21037, 21036, 21033, 21043, 21041.
      iii. For Advanced Parameters: in the dropdown for Annotation File for Collapsing Hits select HIVE ID 30741. Select Collapse Hits By "transcript_id", and Output IDs as "gene_id."
      iv. Click submit.
      v. Once the results are displayed, using the sidebar select View All Available Downloads, and archive "activity-RPKM.csv" and "activity-Hits.csv"
   b. Categorization Table

i. In Microsoft Excel, create a spreadsheet with the first column named "Reference" followed by all of the names of the first columns from the "activity-RPKM.csv" file we just saved.
ii. Then, name the second column of this new spreadsheet "Response" and list below "R" or "NR" according to the known responses recorded. Save this spreadsheet as a .csv file named "Categorization Table."

c. RLDA
i. Back in HIVE, use the upload function to upload the Classification Table we just created to the HIVE Space.
ii. Navigate to Portal > AlgoRLDA (under Classifications). For Matrix File, using the dropdown select the "activity-RPKM.csv" file we just created. Check the box affirming "Samples are first row (instead of first column)."
iii. For Categories File, using the dropdown select the "Categorization Table" file that we just uploaded.
iv. Click submit and the results will load.

4. Results
a. The Contributions table displayed at the top upon completing step "3. c. iv." is an ordered list of the most important predictors.
b. On the same page select Eigenvectors and change the Star from "36" to "20" to clearly see a visualization of the 20 most important predictors.

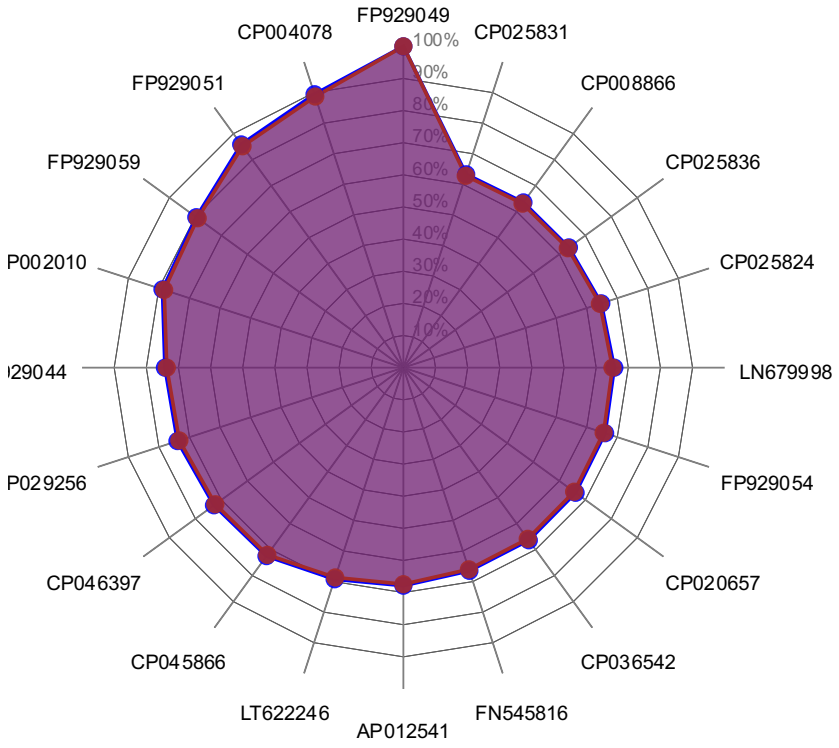

Figure 1: Important Predictors from HIVE RLDA

5. Appendix A: Outside Connections
    a. When cross-referencing these accession numbers with NCBI and including the top predictors from the MATLAB workflow, we can compare the predictors from our computations in HIVE and MATLAB

Rank	From HIVE RLDA	From MATLAB*
1	Roseburia intestinalis, FP929049	Roseburia intestinalis, FP929050
2	Paenibacillus sabinae, CP004078	Bacteroides caesimuris, CP015402
3	Ruminococcus bromii, FP929051	Ruminococcus bromii, FP929051
4	Eubacterium siraeum, FP929059	Clostridium botulinum, CP010520
5	Bifidobacterium longum, CP002010	Pseudomonas tolaasii, CP020369
6	Eubacterium siraeum, FP929044	Enterococcus hirae, CP003504
7	Christensenella minuta, CP029256	Bacteroides dorei, CP008741
8	Bacteroides ovatus, CP046397	Lactobacillus ruminis, CP003032
9	Staphylococcus aureus, CP045866	Eubacterium rectale, FP929043
10	Bacteroides ovatus V975, LT622246	Streptococcus suis, CP000837

Table 1: "MATLAB vs RLDA 1" Prediction Results Compared

    b. From Table 1 above, gather that the top predictor is the same when performing both methods, and the top 3 predictors match 66.6% of the time. However, when looking at the top 10, the predictors only match 30% of the time.

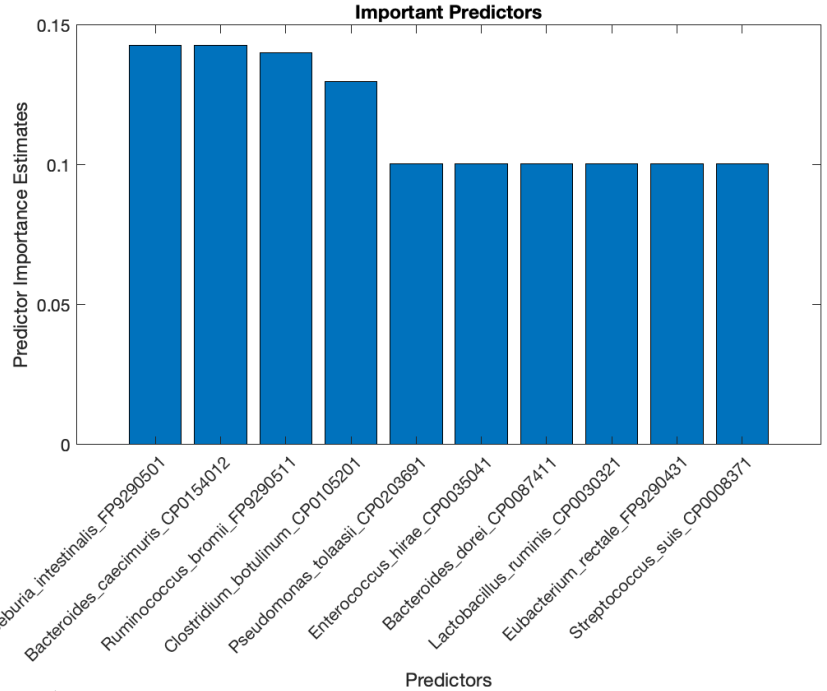

**Figure 2: Important Predictors from MATLAB*

6. Appendix B: Testing HIVE RLDA with a different dataset.
    a. Figure 3 below is a new truncated view of a new dataset that, instead of RPKM values, contains relative abundances of the organisms in the sample that will then become the important predictors.

	Patient_ID	Patient_Before	EffSeizures_After	Akkermansia_muciniphila_CP001071.1	Bacteroides_caccae_CP022412.2	Akkermansia_muciniphila_CP015409.2	E
0	pa03	patient	R	0.000214	0.043905	0.000548	
1	pa04	patient	R	0.003800	0.000300	0.004000	
2	pa05	patient	NR	0.000013	0.128706	0.000007	
3	pa06	patient	R	0.000016	0.251138	0.000034	
4	pa08	patient	NR	0.046365	0.000680	0.106547	
5	pa10	patient	NR	0.081549	0.028902	0.038271	
6	pa11	patient	NR	0.011276	0.000481	0.005397	
7	pa18	patient	R	0.000024	0.076376	0.000010	
8	pa19	patient	NR	0.001712	0.009269	0.004100	
9	pa22	patient	NR	0.055560	0.010666	0.026441	
10	pa23	patient	NR	0.043675	0.045317	0.020533	
11	pa28	patient	R	0.005680	0.245909	0.002886	

12 rows × 190 columns

Figure 3: PA_All.csv

    b. From this dataset, we can create a categorization table that uses the columns "Patient_ID" and "EffSeizures_After." With this new categorization table, we can bring it, and the PA_All dataset, into HIVE RLDA to compute the important predictors with respect to the "EffSeizures_After" column. Below are the eigenvectors of the 10 most important predictors.

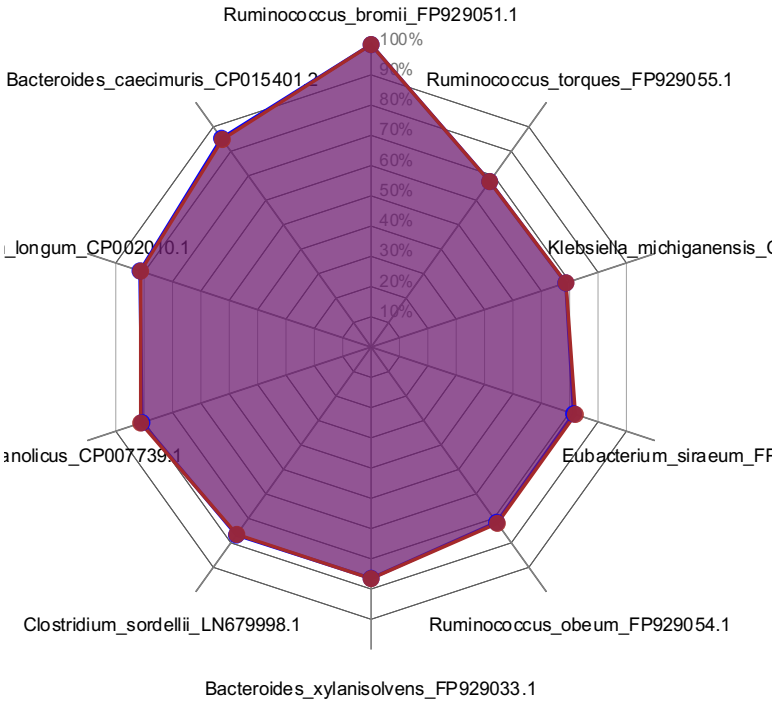

Figure 4: PA_All's Important Predictors from HIVE RLDA

c. Comparing these new important predictors to the ones from MATLAB based on the same dataset (ref. Figure 2) yields the following table:

Rank	From HIVE RLDA	From MATLAB*
1	Ruminococcus bromii, FP929051	Roseburia intestinalis, FP929050
2	Bacteroides caesimuris, CP015402	Bacteroides caesimuris, CP015402
3	Bifidobacterium longum, CP002010	Ruminococcus bromii, FP929051
4	Bacillus methanolicus, CP007739	Clostridium botulinum, CP010520
5	Clostridium sordellii, LN679998	Pseudomonas tolaasii, CP020369
6	Bacteroides xylanisolvens, FP929033	Enterococcus hirae, CP003504
7	Ruminococcus obeum, FP929054	Bacteroides dorei, CP008741
8	Eubacterium siraeum, FP929059	Lactobacillus ruminis, CP003032
9	Klebsiella michiganensis, CP004887	Eubacterium rectale, FP929043
10	Ruminococcus torques, FP929055	Streptococcus suis, CP000837

Table 2: "MATLAB vs RLDA 2" Prediction Results Compared from New Dataset

d. While the top predictors do not align, the second most important predictor matches. And in the top three predictors, two organisms appear in both methods. Overall, the top ten predictors only share one directly ranked organism, with there only being two organisms that appear in both important predictor sets. The correlations of results with the new dataset between these two methods are weaker than the previous comparison correlations.

7. Appendix C: Comparing Similar Predictors
    a. If we take a look at the Top 10, 25, 50 and 100 similar predictors between both MATLAB and RLDA from Hexagon RPMK (MATLAB vs RLDA 1) and MATLAB and RLDA from abundance values, as seen in Step 6. a. (MATLAB vs RLDA 2), the following graph can be generated. It is important to note that the values have been normalized to the sample size of predictors.

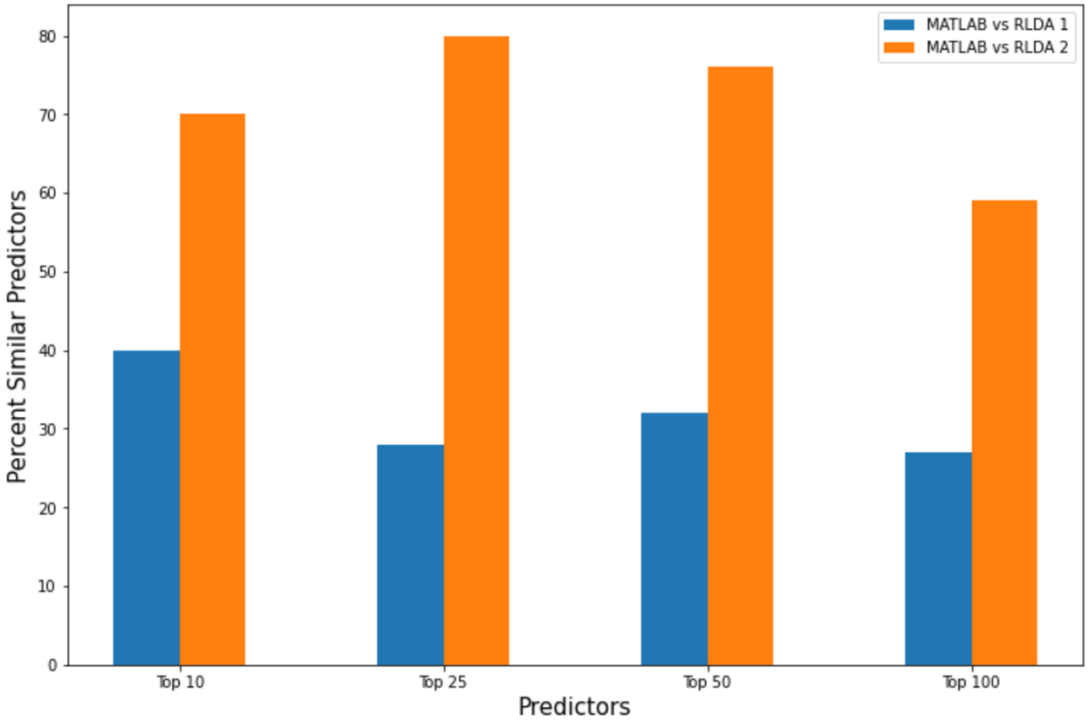

Figure 5: Percent of Similar Predictors per Comparison

    b. From this, it can be gathered that comparing RLDA important predictors gathered from relative abundances of organisms rather than RPKM values to MATLAB generated important predictors is the better of the two comparisons to make. Furthermore, conclude from this analysis that it is prudent to evaluate at least the top 25 predictors in order to identify overlap between MATLAB and RLDA as implemented in HIVE.

References:

1. Smith AD, Foss ED, Zhang I, Hastie JL, Giordano NP, Gasparyan L, VinhNguyen LP, Schubert AM, Prasad D, McMichael HL, Sun J, Beger RD, Simonyan V, Cowley SC, Carlson PE Jr. Microbiota of MR1 deficient mice confer resistance against Clostridium difficile infection. PLoS One. 2019 Sep 27;14(9):e0223025. doi: 10.1371/journal.pone.0223025. PMID: 31560732; PMCID: PMC6764671.