**Gut-Microbiome Analysis: Finding Important Predictors within Responding and Non-Responding Patients with respect to a Ketogenic Diet – Justification for multi-pronged analysis pipeline**

**5/27/2021**

**James Ziegler1, Stephanie Singleton1, Raja Mazumder1**

*1HIVE Lab, Department of Biochemistry and Molecular Medicine, The George Washington University, Washington DC, 20052*

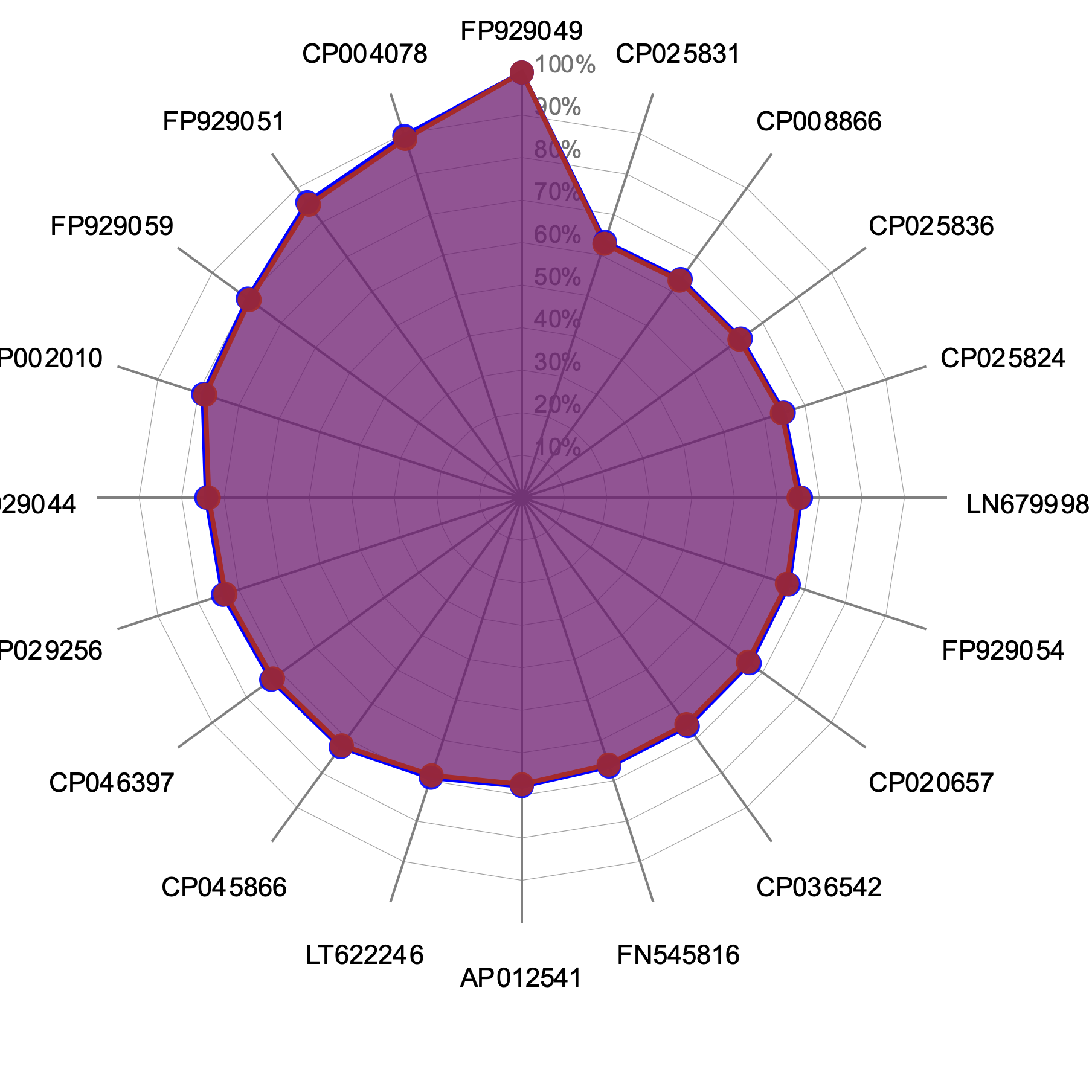
1. **Summary**

Here we describe methods to identify which organisms are important predictors of a patient’s associated response to a ketogenic diet (KD), where response values are designated as either 50% or greater reduction in seizure frequency (Responder (R)), or less than 50% reduction in seizure frequency (Non-respond (NR)). The data used can be described as metagenomics samples collected from 12 patients with Epilepsy before (PA; sample data 1) undergoing a KD for 3 months, and after (PB; sample data 2)1. Sample data 1 was paired with known response values collected at the completion of the study (R or NR), allowing sophisticated tools such as Machine Learning and Regularized Linear Discriminant Analysis2 (RLDA) to identify important predictors associated with the response value and the gut microbiome profile of the patient. Here, we implement RLDA in HIVE3, where we demonstrate that it is prudent to evaluate at least the top 25 predictors to identify overlap between MATLAB and RLDA as implemented in HIVE.

1. **Objectives**

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

1. **Methods**
   1. RPKM Table
      1. In HIVE navigate to Portal > Alignment Comparator (under Classifications).
      2. 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.
      3. 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.”
      4. Click submit.
      5. Once the results are displayed, using the sidebar select View All Available Downloads, and archive “activity-RPKM.csv” and “activity-Hits.csv”
   2. Categorization Table
      1. 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.
      2. 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.”
   3. RLDA
      1. Back in HIVE, use the upload function to upload the Classification Table we just created to the HIVE Space.
      2. 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).”
      3. For Categories File, using the dropdown select the “Categorization Table” file that we just uploaded.
      4. Click submit and the results will load.
2. **Results**
   1. The Contributions table displayed at the top upon completing step “3. c. iv.” is an ordered list of the most important predictors.
   2. 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

1. **Appendix A: Outside Connections**

|  |  |  |
| --- | --- | --- |
| 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

* + - * 1. 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
        2. 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.

Chart, bar chart

Description automatically generated

***Figure 2***: Important Predictors from MATLAB

1. **Appendix B: Testing HIVE RLDA with a different dataset.**
   1. 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.

Table

Description automatically generated

**Figure 3:** PA\_All.csv

* 1. 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.

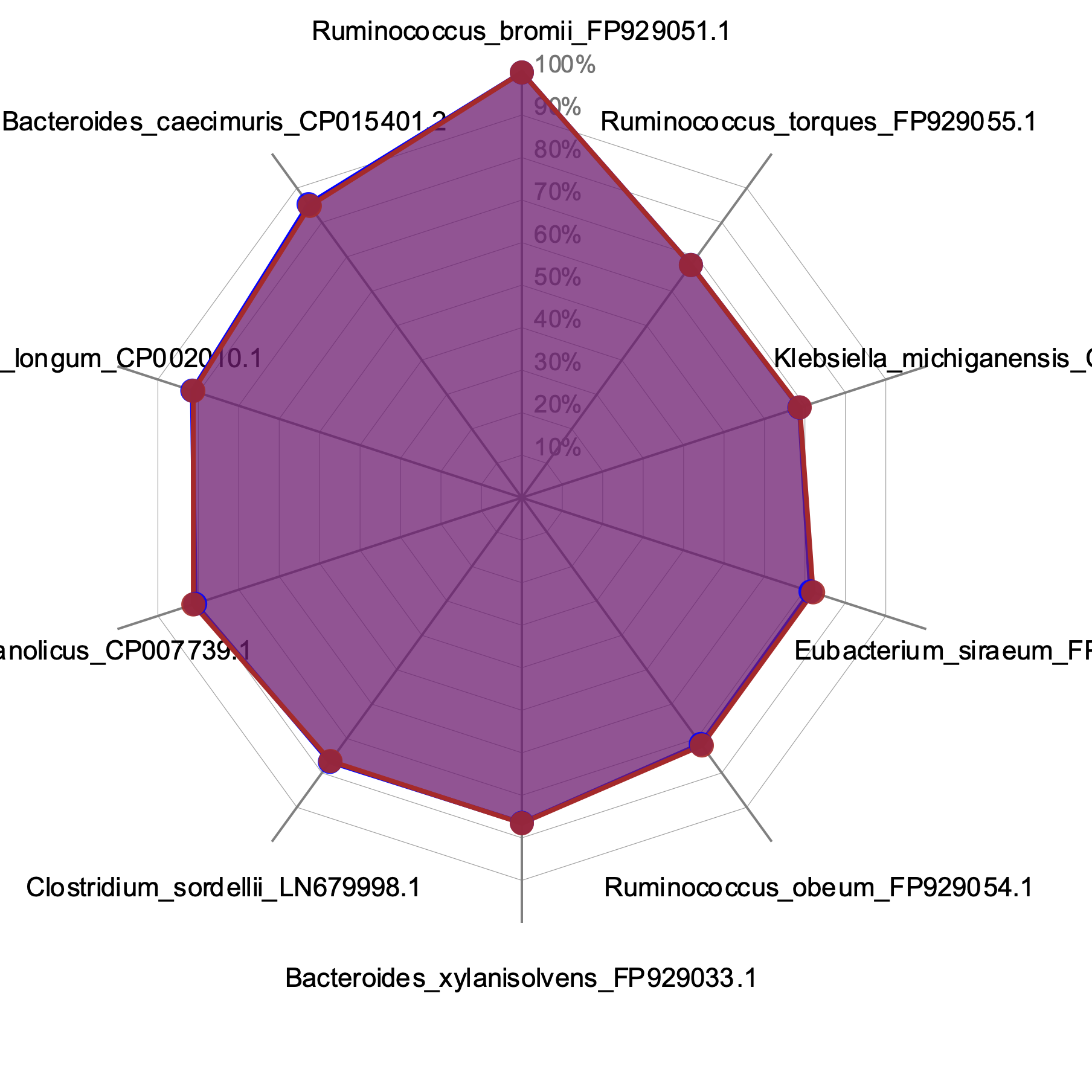


Figure 4: PA\_All’s Important Predictors from HIVE RLDA

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

**Appendix C: Comparing Similar Predictors**

|  |  |  |
| --- | --- | --- |
| 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

* 1. 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.
  2. 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**

Chart, bar chart

Description automatically generated

**Figure 5:** Percent of Similar Predictors per Comparison

1. Lindefeldt, M. et al. The ketogenic diet influences taxonomic and functional composition of the gut microbiota in children with severe epilepsy. *NPJ Biofilms Microbiomes* **5**, 5 (2019).

2. Smith, A.D. et al. Microbiota of MR1 deficient mice confer resistance against Clostridium difficile infection. *PLoS One* **14**, e0223025 (2019).

3. Santana-Quintero, L., Dingerdissen, H., Thierry-Mieg, J., Mazumder, R. & Simonyan, V. HIVE-hexagon: high-performance, parallelized sequence alignment for next-generation sequencing data analysis. *PLoS One* **9**, e99033 (2014).