



**HAL**  
open science

# FEATURE SELECTION APPLIED TO MICROBIOME FOR DRUG DISCOVERY

David Rojas-Velazquez, Sarah Kidwai, Lara Cerezin, Luciënne de Vries, Kosta Besermenji, Paula Perez-Pardo, Alejandro Lopez-Rincon

► **To cite this version:**

David Rojas-Velazquez, Sarah Kidwai, Lara Cerezin, Luciënne de Vries, Kosta Besermenji, et al.. FEATURE SELECTION APPLIED TO MICROBIOME FOR DRUG DISCOVERY. CMBBE 2023 - 18th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering (CMBBE 2023), May 2023, Paris, France. hal-04097608

**HAL Id: hal-04097608**

**<https://hal.science/hal-04097608>**

Submitted on 15 May 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# FEATURE SELECTION APPLIED TO MICROBIOME FOR DRUG DISCOVERY

David Rojas-Velazquez (1,2), Sarah Kidwai (1), Lara Cerezin(1), Luciënne de Vries(1), Kosta Besermenji(1), Paula Perez-Pardo(1), Alejandro Lopez-Rincon (1,2)

1. Utrecht University, Pharmacology, Netherlands; 2. Julius Center UMC Utrecht, Netherlands

## 1. Introduction

Inflammatory bowel disease (IBD) is a term that describes conditions characterized by chronic inflammation of the gastrointestinal (GI) tract caused by various factors: abnormal gut microbiota, immune response dysregulation, environmental changes and gene variants. There are two types of IBD: Crohn's disease (CD) and ulcerative colitis (UC). Currently, there is no treatment for IBD, but it can be managed by using aminosalicylates, immunosuppressants or biologics<sup>1</sup>.

## 2. Materials and Methods

The data used is from *Alam et al* [1], where fecal samples were collected from 30 individuals (20 IBD and 10 healthy volunteers) with 16S rRNA taxonomic profiling. Raw data was analysed using the DADA2 pipeline [2] and the Recursive Ensemble Feature Selection (REFS), and we compared with the original results. REFS is a method to discover biomarkers, the ensemble for the feature selection phase is composed by 8 classifiers from the sci-kit learn toolbox [3]: Stochastic Gradient Descent, Support Vector Machine classifier, gradient boosting, random forest, logistic regression, passive aggressive classifier, ridge classifier and bagging [4,5]. Once the features were selected, are validated using 5 different classifiers from the sci-kit learn toolbox [3] not part from the ensemble.

## 3. Results

REFS selected 5 sequences (features) from the original 3226, with an AUC of 0.92, considered as excellent in medical diagnostics [6]. The resulting taxa are 5 out 3 at genus level: *Lactobacillus* (F1), *UCG-002* (F2) and *Fusicatenibacter* (F5), and 2 *Lachnospirales* that will require more study in tools such as BLAST<sup>2</sup>.

## 4. Discussion and Conclusions

As shown in Figure 2, the use of ML with REFS give us better results and closer to reality. E.g. the first feature selected was genus *Lactobacillus*



Figure 1: 10 runs of the REFS algorithm, with best answer at 5 taxa (vertical red line).

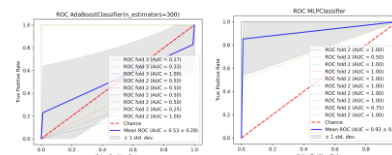


Figure 2: ROC Comparison between the original results from [1] (left) and the results using REFS (right).

and it was shown to be underexpressed in IBD patients. *Lactobacillus* is a component of lactic acid bacteria, a bacteria group described by the formation of lactic acid as a main end product of carbohydrate metabolism [7]. *Lactobacilli* are a major part of the commensal microbial flora of small and large intestine in humans and animals, and are often used as probiotics [8]. An overview in [9] described the use of VSL #3 which consists of 8 bacterial strains, 4 of those are *Lactobacillus* strains. It has been shown that VSL #3 was able to induce a significant increase in protective bacterial strains and can induce remission in patients with mild to moderate UC [9], this could be a first step towards a treatment.

## 5. References

1. Alam MT et al., Gut pathogens. 2020;12(1):1-8 (2020).
2. Callahan B et al., Nature methods. 2016;13(7):581-583 (2016).
3. Pedregosa F et al., The Journal of machine Learning research. 2011;12:2825-2830 (2011).
4. Lopez-Rincon A et al., BMC bioinformatics. 2019;20(1):1-17 (2019).
5. Lopez-Rincon A et al., BMC bioinformatics. 2020;12(7):785 (2020).
6. Šimundić AM, ejfccc. 2009;19(4):203 (2009).
7. Tannock GW, Applied and Environmental Microbiology. 2004;70(6):3189-94 (2004).
8. Christensen HR, The Journal of Immunology. 2002;168(1):171-8 (2002).
9. Cheng FS, World Journal of Clinical Cases. 2020; 8(8):1361-84 (2020).

<sup>1</sup> <https://www.nhs.uk/conditions/inflammatory-bowel-disease/>

<sup>2</sup> <https://blast.ncbi.nlm.nih.gov/Blast.cgi>