PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Are metagenomics data sufficiently informative for potential non-invasive diagnosis of inflammatory bowel disease status — Outcomes of the crowdsourced sbv IMPROVER MEDIC challenge

Carine Poussin, Lusine Khachatryan, Yang Xiang, Adrian Stan, James Battey, Giuseppe Lo Sasso, Stephanie Boue, Nicolas Sierro, Nikolai V. Ivanov, Manuel C Peitsch, Julia Hoeng

[1] *PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland*

INTRODUCTION AND OBJECTIVES

A growing body of evidence links gut microbiota changes with inflammatory bowel disease (IBD), raising the question of the potential benefit of exploiting metagenomics data for non-invasive IBD diagnostics. Open between September 2019 and March 2020, the sbv IMPROVER Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge (MEDIC) investigated computational metagenomics methods for discriminating IBD and non-IBD subjects. For developing and applying models for classifying metagenomics fecal samples, participants were offered the option to start with raw (sub-challenge 1, SC1) or taxonomy- and pathway-based processed (sub-challenge 2, SC2) independent training and test metagenomics data from IBD and non-IBD subjects. We have received and scored a total of 81 anonymized submissions. The results show that many participants' predictions performed better than random predictions for classifying IBD vs. non-IBD, ulcerative colitis (UC) vs. non-IBD, and Crohn's disease (CD) vs. non-IBD. However, discrimination of UC and CD remains challenging, with very few submissions reaching the level of significance. Following the challenge, we are conducting an analysis of class predictions and metagenomics features across the teams, including evaluation of the computational methods used to solve the problem. These results will be openly shared with the scientific community to help advance research in the field of IBD.

MATERIALS AND METHODS

The sbv IMPROVER MEDIC in 2019–2020 aimed to **explore the diagnostic potential of metagenomics sequencing data** to:

(1) Differentiate IBD and non-IBD subjects
(2) Within the IBD group, discriminate between UC and CD subjects

Scientific Questions

- Which predictive computational model provides the most accurate classification?
- What does the most discriminative metagenomics signature(s) tell us?

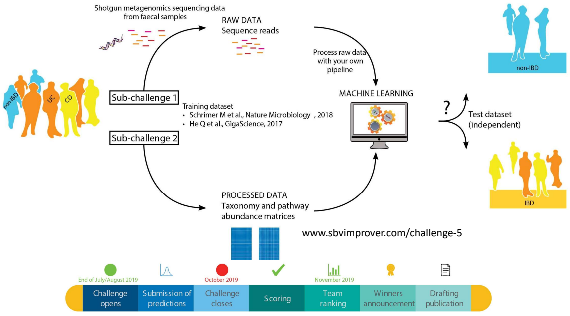

Shotgun metagenomics sequencing data from faecal samples
RAW DATA
Sequence reads
Process raw data with your own pipeline
MACHINE LEARNING
non-IBD
UC
CD
Sub-challenge 1
Sub-challenge 2
Training dataset
• Schrimer M et al., Nature Microbiology , 2018
• He Q et al., GigaScience, 2017
?
Test dataset (independent)
non-IBD
IBD
PROCESSED DATA
Taxonomy and pathway abundance matrices
www.sbvimprover.com/challenge-5
End of July/August 2019
October 2019
November 2019
Challenge opens
Submission of predictions
Challenge closes
Scoring
Team ranking
Winners announcement
Drafting publication

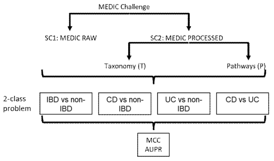
Figure 1. *Overview of the sbv IMPROVER MEDIC challenge.*

Challenge Data

Participants were asked to leverage publicly available metagenomics data from Schirmer et al. [1,2] and He et al. [3] for training their classification models, either starting from raw sequencing data to apply their own metagenomics pipeline for SC1 or by using the taxonomy- and pathway-based relative abundances data provided by the challenge organizers for SC2. Participants applied their trained models on an independent metagenomic dataset to predict the class labels of the subjects from whom the analyzed samples were derived.

Scoring

Qualified submissions (fulfilled challenge requirements) were anonymized and scored by comparing the predicted class labels with the true class labels (Gold standard) by using specific and complementary predefined metrics, including the area under the precision recall (AUPR, [4]) and the Matthews correlation coefficient (MCC, [5]). Scores were aggregated and ranked from the best (lowest rank) to the worst (highest rank) performance for SC1 and SC2, separately. The scoring results and final list of best performing teams have been reviewed and approved by an external and independent scoring review panel of experts.



$$R_{problem} = \frac{R_{problem}^{AUPR} + R_{problem}^{MCC}}{2} \quad (1)$$

$$Weighted\ Sum\ of\ Ranks_{\ SC1} = R_{IBD\ vs\ non-IBD} + R_{CD\ vs\ non-I} \quad + R_{UC\ vs\ non-IBD} + 2 \times R_{CD\ vs\ UC} \quad (2)$$

$$Weighted\ Sum\ of\ Ranks_{\ SC2} = \frac{1}{2} \times \{(R_{IBD\ vs\ non-I} \quad + R_{CD\ vs\ non-IBD} + R_{UC\ vs\ non-I} \quad + 2 \times R_{CD\ vs\ UC})_T + (R_{IBD\ vs\ non-I} \quad + R_{CD\ vs\ non-I} \quad + R_{UC\ vs\ non-I} \quad + 2 \times R_{CD\ vs\ UC})_P\} \quad (3)$$

Figure 2. *Scoring of the MEDIC challenge. Participants were requested to classify samples given the four 2-class problems posed for SC1 and SC2. The AUPR and MCC were used as complementary metrics and computed for each submission. To evaluate whether predictions were better than random, the AUPR and MCC scores were compared with the 95th percentile of AUPR and MCC score distributions calculated from 10,000 random predictions, respectively. When a participants' score was lower than the 95th percentile score, the score was set to the value of the 95th percentile score and considered to be non-significant. For each 2-class problem and metric, all submission scores were ranked. Ranks (R) associated with AUPR and MCC scores were averaged for each submission (1). Ranks were aggregated across the four 2-class problems by calculating a weighted sum of ranks (2 for SC1 and 3 for SC2). For SC2, the weighted sum of ranks included the ranks obtained when predicting class labels by using taxonomy (T) and pathway (P) relative abundances data (3). The best performing teams obtained the lowest weighted sum of ranks.*

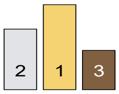
BEST PERFORMERS

Sub-challenge 1 team leader

1. Artem Ivanov (1 member) – ITMO University, Russia
2. Garrett Graham (1 member) - Georgetown University Medical Center, USA
3. Mario Rosario Guarracino (6 members) – HPC and Networking Institute CNRS, Italy

Sub-challenge 2 team leader

1. Artem Ivanov (1 member) – ITMO University, Russia
2. Enrico Glaab (1 member) – University of Luxemburg, Luxemburg
3. Barbara Di Camillo (7 members) – University of Padova, Italy



REFERENCES

[1] Schirmer M (2018) and [2] https://www.ibdmdb.org/
[3] He Q (2017)
[4] Saito T and Rehmsmeier M (2015)
[5] Matthews BW (1975)

ACKNOWLEDGEMENTS

Scoring Review Panel
- Dr Laurent FALQUET - University of Fribourg, Fribourg, Switzerland
- Prof. Dr. Prashantha Karunakar , PES University, Bangalore, India
Project management, Finance: Gill Den Hartog, Cédric Montandon

RESULTS

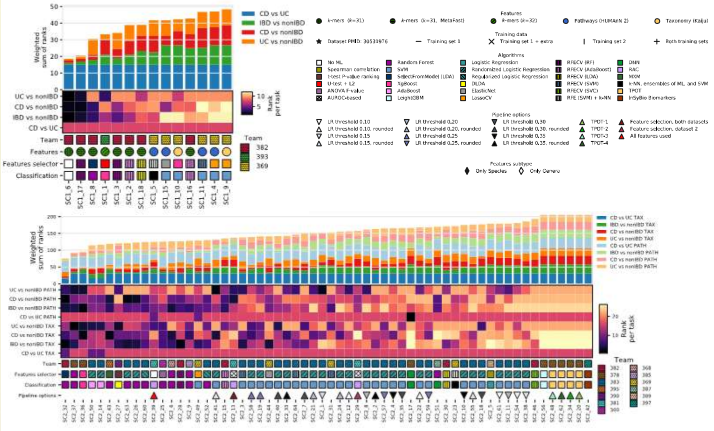


Figure 3. *Weighted sum of ranks and final ranking of the submissions for SC1 and SC2.The bar plot represents the weighted sum of ranks, and the colors within bars show the contribution of each task to the final sum of ranks. The heatmap represents the averaged ranks per classification task. The submissions are ordered from the lowest (best) to highest (worst) weighted sum of ranks. The color code at the bottom of the heatmap shows the association between individual teams and submissions. The machine learning approaches used by the participants were: linear discriminant analysis (LDA), random forest (RF), support vector machine (SVM), k-nearest neighbours (kNN), support vector classifier (SVC), deep neural networks (DNN), and logistic regression (LR).*

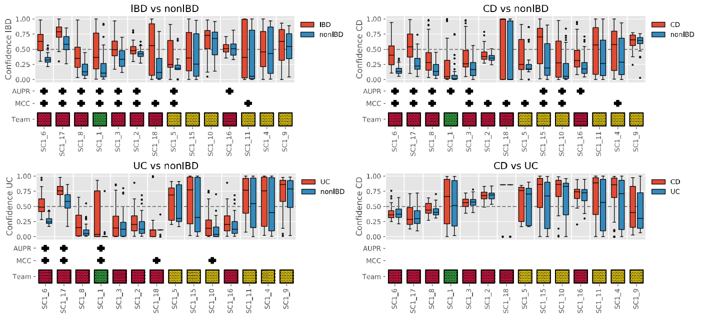


Figure 4. *Boxplots of confidence values that a sample belongs to class 1 (red) or class 2 (blue) for each submission and 2-class problem of SC1. Submissions are colored by team at the bottom and ordered from the best to worst overall performance.*

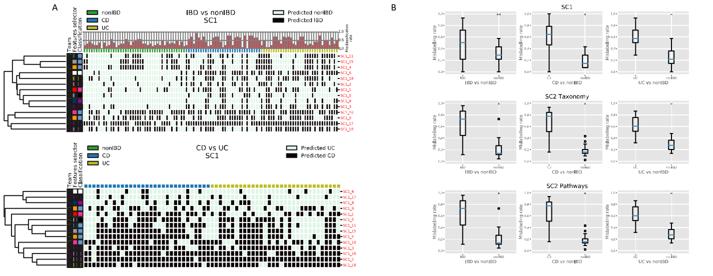


Figure 5. *Sample misclassification. (A) Most frequently misclassified samples are highlighted in heatmaps for IBD vs. non-IBD or CD vs. UC, provided as examples; (B) Boxplots of the misclassification rate per class for each 2-class problems for SC1 and SC2.*

CONCLUSIONS

- In total, **81 submissions** were received for the sbv IMPROVER MEDIC challenge from **participants worldwide.**

Initial post-challenge analysis results show that:

- **Metagenomics data generated from fecal samples are sufficiently informative to discriminate non-IBD and IBD status.**
- However, within the IBD group, **discriminating UC and CD remains challenging.**
- Classification using **k-mers-based features showed a better performance than classification using the mapping-based features** (taxonomy and pathways) provided for SC2.
- The **type of algorithms that performed the best varies depending on the 2-class problem.**
- On the basis of overall performance, **tree-based classification methods demonstrated the best performance in both sub-challenges.**
- **IBD samples were more frequently misclassified than non-IBD samples.**

Competing financial interest: The research described in this poster was sponsored by Philip Morris International