Systematic Evaluation of Data Processing Pipelines and Machine Learning Classifiers for Shotgun Metagenomics-based Diagnosis of Crohn's Disease and Colorectal Cancer

Sungho Lee[1] and Insuk Lee[1*]

[1] Department of Biotechnology, College of Life Science & Biotechnology, Yonsei University, Seoul 03722, Republic of Korea

[*] Correspondence: insuklee@yonsei.ac.kr

Introduction

Ever since humanity recognized the physiopathological correlation between the intestinal microbes and the host, there have been increasing efforts to utilize the gut microbiome as a predictive tool for human diseases, such as Crohn's disease (CD) and colorectal cancer (CRC). Advances in sequencing technologies such as shotgun metagenomics and the application of machine learning (ML), a breakthrough for various data-driven approaches, have greatly accelerated these endeavors. Despite this, there is still a lack of consensus on the most suitable data processing methods and ML algorithms for shotgun metagenomics-based disease diagnosis, and the fitness and relative performance of existing computational tools should be benchmarked further.

Results

We performed a systematic benchmark of shotgun metagenomics-based disease diagnostic ML pipelines using 2,553 public fecal metagenomic samples collected from 21 case-control studies of CD and CRC, focusing on two essential prerequisites for successful clinical application: classification performance and generalizability to previously unseen datasets. For this purpose, we systematically evaluated the effects of 12 microbiome-based features, 5 batch correction methods, 7 normalization techniques, and 9 ML models on the disease classification capability by summarizing the binary classification performance of the entire 5,184 combinations under the 20-fold leave-one-dataset-out cross-validation scheme, resulting in a total of 1,658,880 ML model instances.

We confirmed that the highest level of performance was attained by non-linear ensemble ML models, which were trained utilizing gut-specific taxonomic features that had undergone compositional transformation. The benchmarking outcomes also revealed that the reduction of variance linked to biological context during the batch correction procedure resulted in a diminished overall performance, regardless of the extent of batch mixing. Based on these insights, we orchestrated the optimal ML pipelines for CD and CRC and demonstrated the decent performance of trained diagnostic models using holdout test cohorts from multiple countries.

Finally, we outlined the behavior and decision-making process of trained ML models from the extracted feature importance and identified that each diagnostic model learned disease-specific gut microbial signatures. In CD, impacts of previously unknown species were discovered in addition to some well-known species such as *Klebsiella pneumoniae*, while in CRC, previously recognized disease-associated microbes such as *Fusobacterium nucleatum* subspecies, *Prevotella intermedia, Allisonella pneumosintes*, and *Gemella morbillorum* were pinpointed as having a strong influence on model decisions.

Conclusion

Overall, the benchmark provides a standardized approach and practical suggestions for establishing robust disease diagnostic ML models from multi-cohort shotgun metagenomic datasets.