The Importance of Taxonomic Classification Software and Machine Learning Algorithms for the Prediction of Colorectal Cancer

Mølvang Dall, Sebastian; Yssing Michaelsen, Thomas; Albertsen, Mads

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Mølvang Dall, S., Yssing Michaelsen, T., & Albertsen, M. (2021). *The Importance of Taxonomic Classification Software and Machine Learning Algorithms for the Prediction of Colorectal Cancer*. Poster presented at Danish Microbiological Society congress 2021, Copenhagen, Denmark.

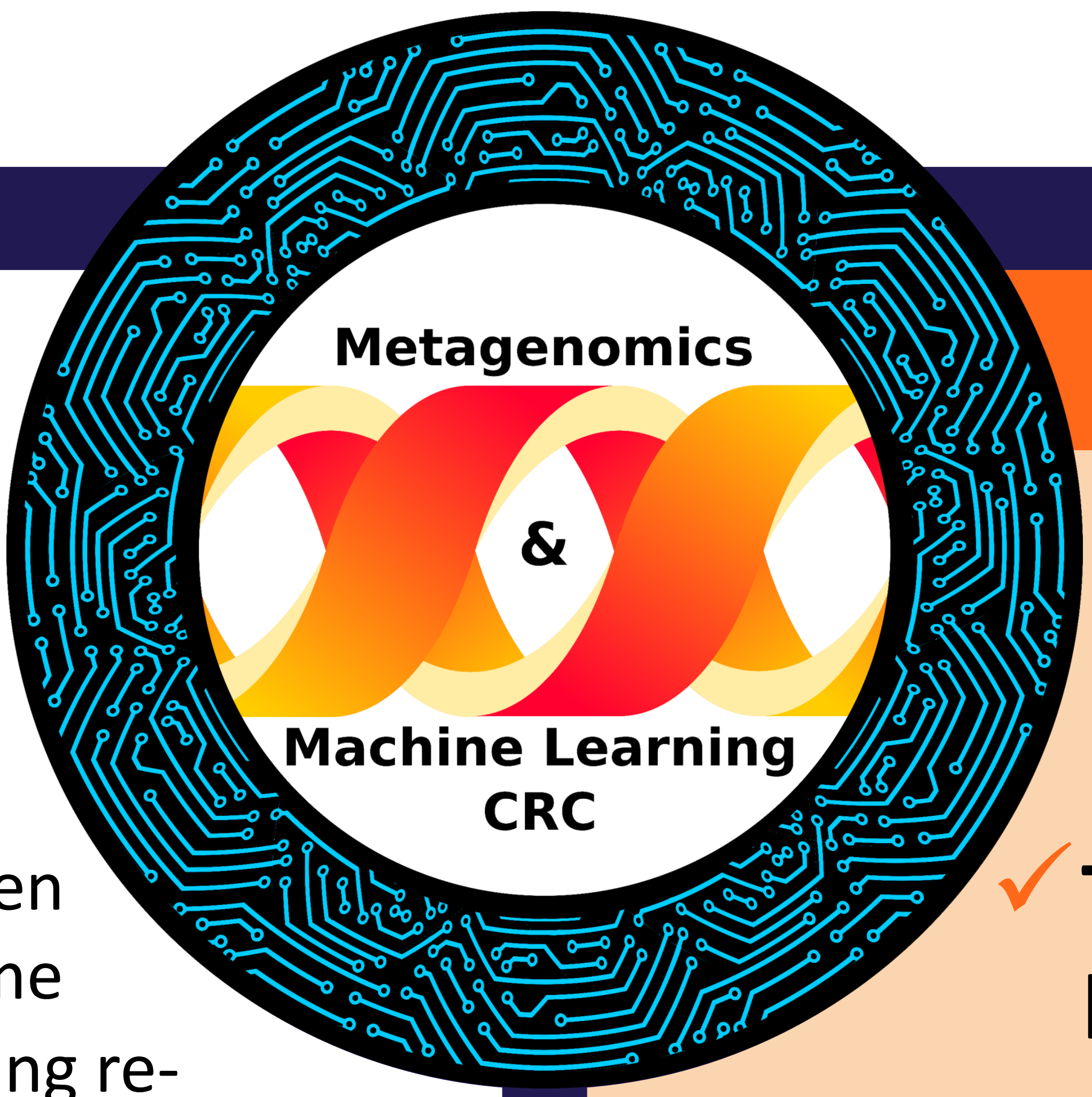
The Importance of Taxonomic Classification Software and Machine Learning Algorithms for the Prediction of Colorectal Cancer

AALBORG UNIVERSITY

Introduction

Colorectal Cancer (CRC) represents a rising global burden ranking 3rd and 2nd in terms of cancer incidence and mortality, respectively.

In recent years strong associations have been revealed between the human gut microbiome and CRC. Metagenomics and machine learning represent a new diagnostic tool for CRC, however the effect of taxonomic classification software and machine learning algorithms on CRC prediction have not been investigated.



Findings

✓ **Kraken2** output results in the average **best CRC prediction.**

✓ **Tree-based** algorithms produce the best models for CRC prediction.

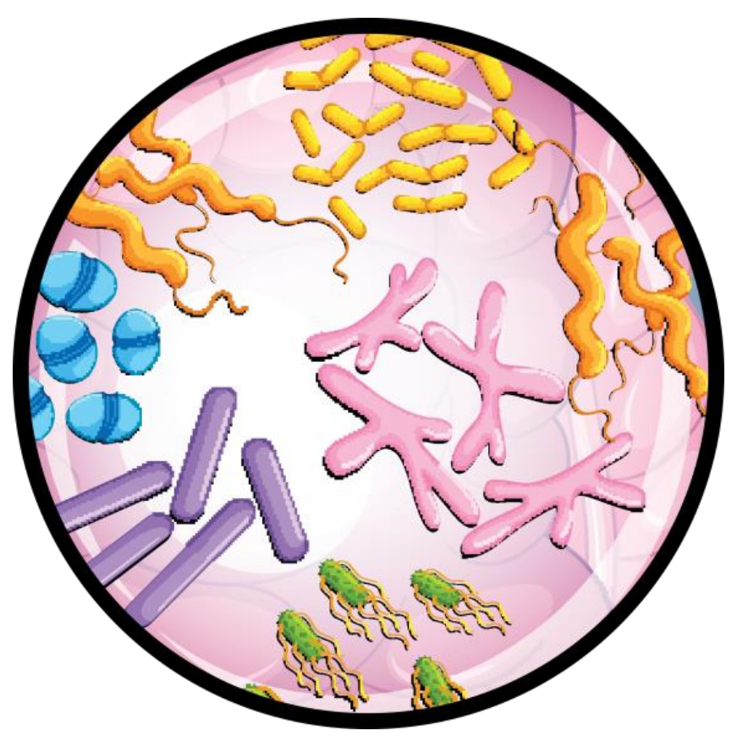
✓ **CRC prediction** can be achieved with just 100,000 reads with **Kraken2.**

Methods

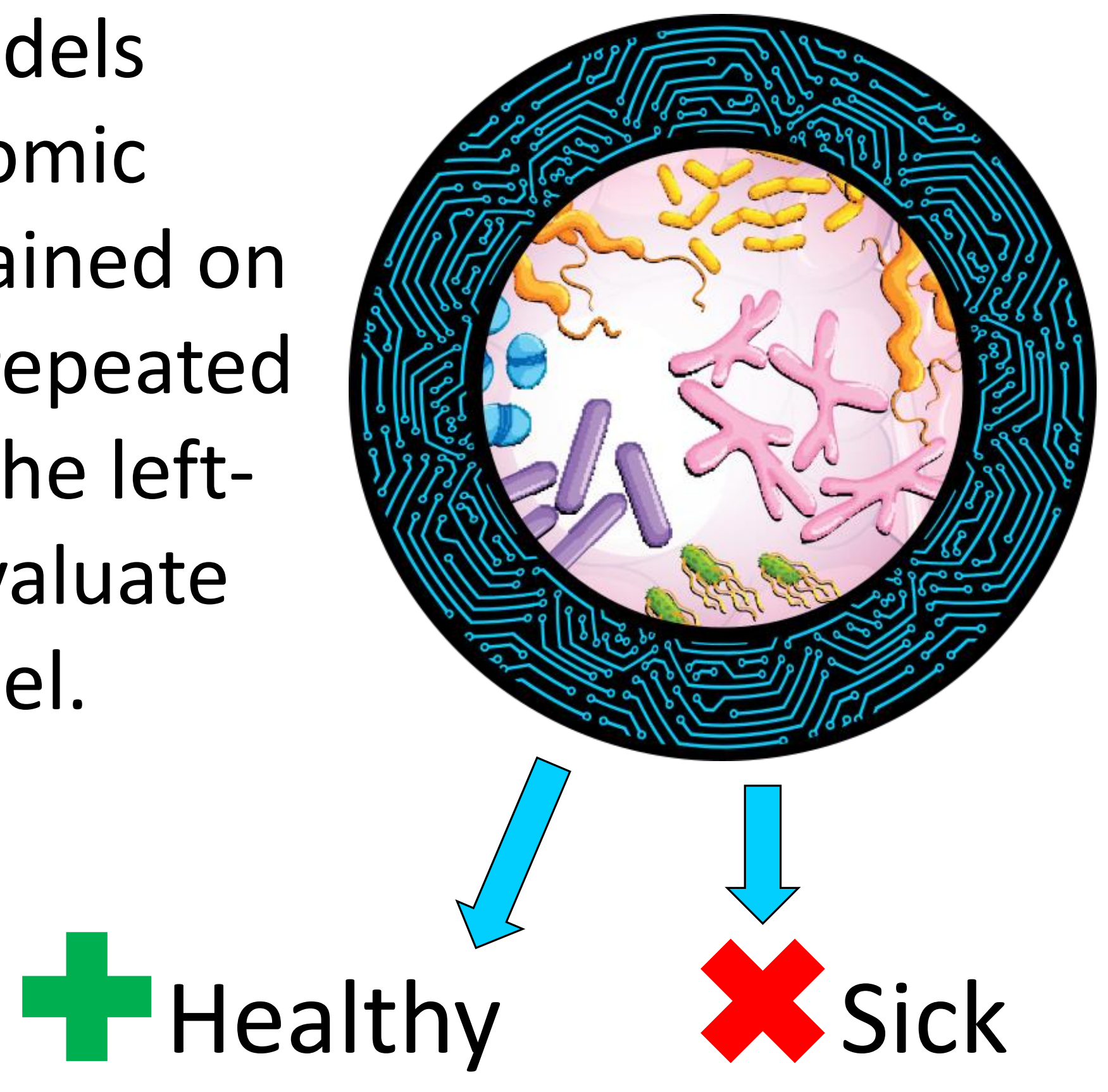
Study (Year)	Country	CRC	CTR
Zeller, G (2014)	FR (114) DE (38)	91	61
Yu, J (2015)	CN	74	54
Feng, Q (2015)	AT	46	63
Vogtmann, E (2016)	US	52	52
Thomas, A (2019)	IT	61	52
Yachida, S (2019)	JP	258	251
Wirbel, J (2019)	DE	22	60
Gupta, A (2019)	IN	30	30

1. **DNA sequences** from fecal were downloaded from the studies. Sequences went through quality check and human DNA was removed.

2. **Kaiju, Kraken2, and MetaPhlAn3** were used to make taxonomic profiles.



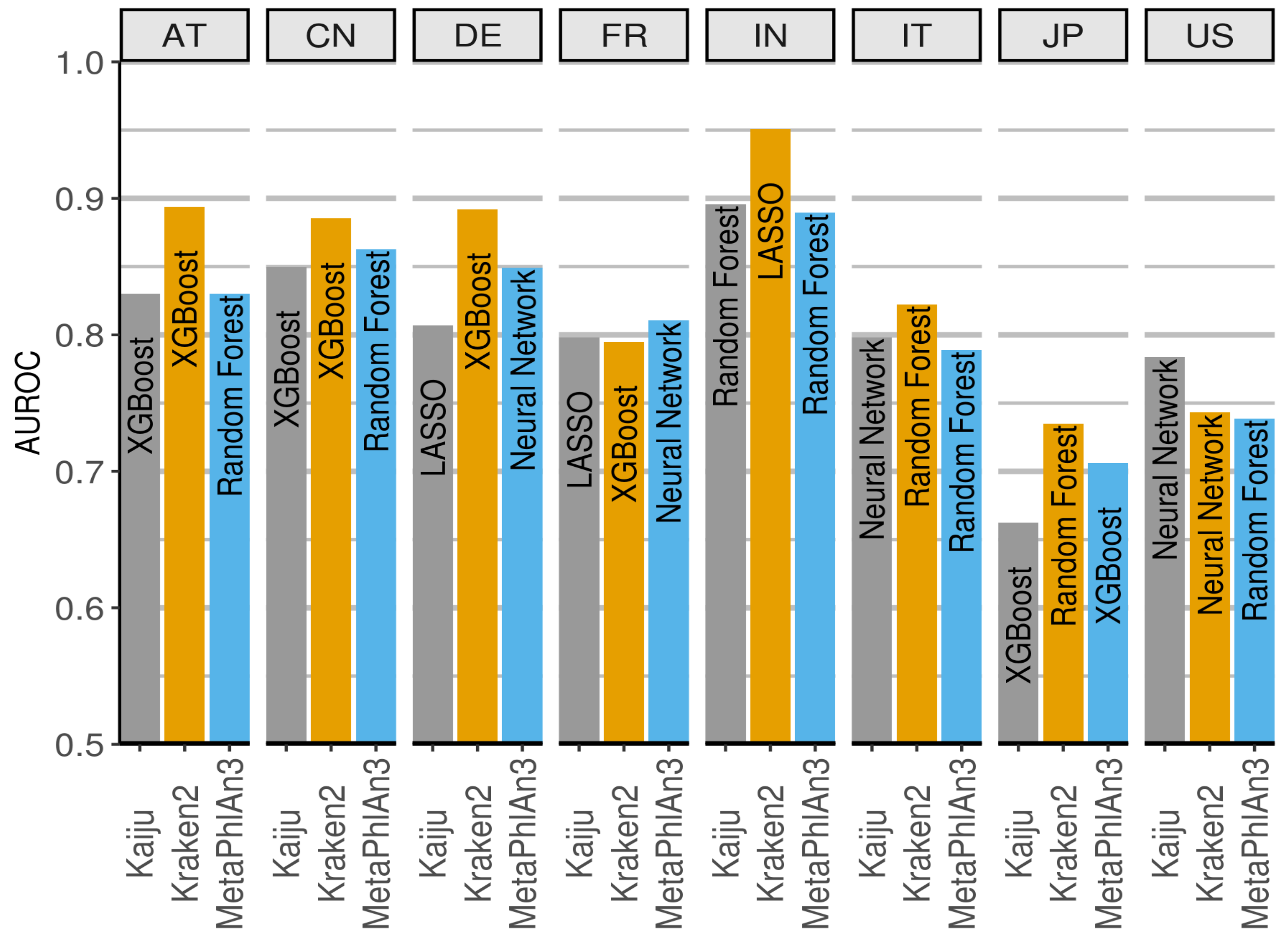
3. **Logistic LASSO regression, Random Forest, XGBoost, and MLP neural network** models were build on the taxonomic profiles. Models were trained on 7 out of 8 studies using repeated 5-fold cross-validation. The left-out study was used to evaluate performance of the model.



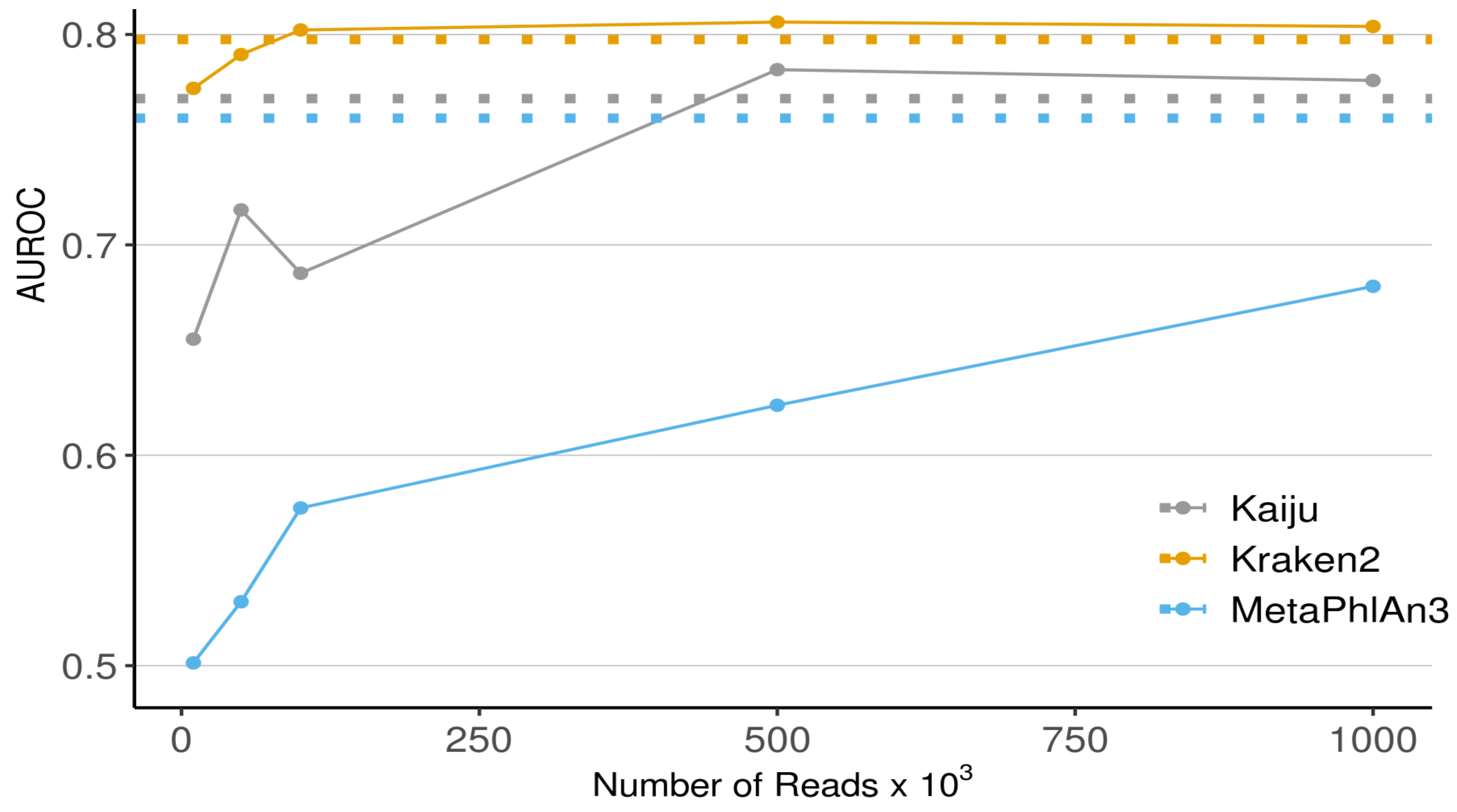
Sebastian Mølvang Dall

✉ semoda@bio.aau.dk

Results



- The best model produced was built with the Kraken2 output six out of eight times.
- The best model for a given study was most often a XGBoost or random forest model (Tree-based algorithm).



- Average performance for LASSO models built on subsampled samples. Full CRC detection could be achieved with 100,000 reads with Kraken2.
- MetaPhlan3 is more affected by subsampling than Kaiju and Kraken2.