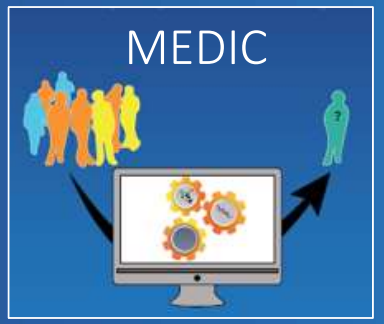
The sbvIMPROVER Metagenomics Diagnostics for Inflammatory Bowel Disease Challenge:

Results and lessons learned



sbv IMPROVER
SYSTEMS BIOLOGY VERIFICATION

www.sbvimprover.com

July 7 2020

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Outline

- What is sbv IMPROVER ?
- Background: Inflammatory Bowel Disease (IBD) and microbiome
- The Metagenomics Diagnosis for Inflammatory Bowel Disease Challenge
- Scoring
- First results
- Conclusions and future plans

WHAT IS SBV IMPROVER ?

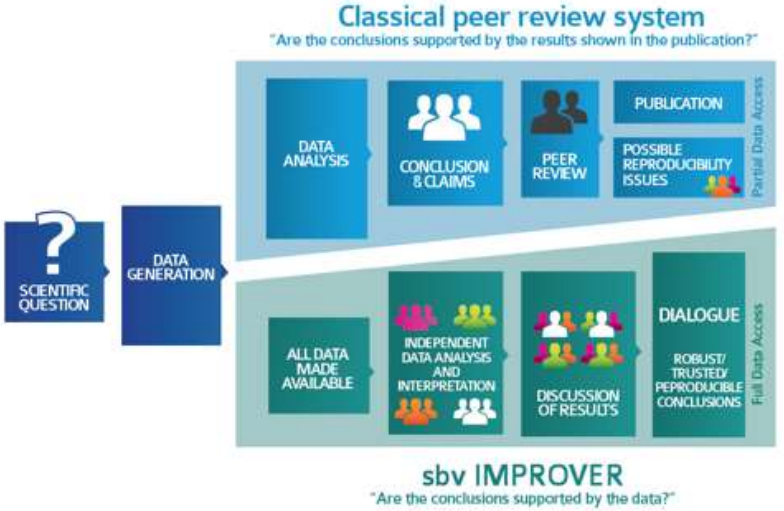
sbv IMPROVER

sbv IMPROVER stands for **S**ystems **B**iology **V**erification combined with **I**ndustrial **M**ethodology for **Pro**cess **Ve**rification in **R**esearch.

This approach aims to provide a measure of quality control in industrial research and development by verifying the methods used. It is complementary to the classical peer-review system.

Double-blind performance assessment to address the concern of self-assessment trap (*Norel R, Molecular Systems Biology, 2011*)

The sbv IMPROVER project is a collaborative effort led and funded by PMI Research and Development.




BIOINFORMATICS REVIEW

Systems biology

Industrial methodology for process verification in research (IMPROVER): toward systems biology verification


COMMENTARY

Verification of systems biology research in the age of collaborative competition

Nature Biotechnology 2011 Sep 8;29(9):811-5
Bioinformatics 2012 28(9):1193-1201

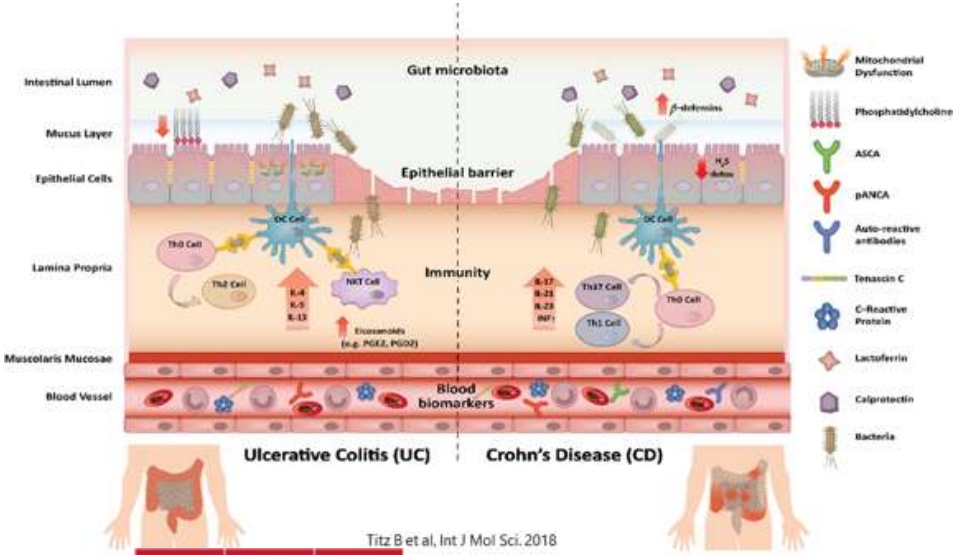
IBD AND MICROBIOME ?

Inflammatory Bowel Disease (IBD)

- Inflammatory bowel diseases (Crohn's disease and ulcerative colitis) are chronic idiopathic disorders that cause inflammation of the gastrointestinal tract.
- Historical and epidemiological data from the last century suggest that the emergence of IBD followed the industrialization and westernization of society.
- Various studies have suggested a strong connection between these diseases and the composition of gastrointestinal tract microflora.

Kaplan G.G et al. 2015

Nature Reviews | Gastroenterology & Hepatology



Titz B et al, Int J Mol Sci. 2018

	Ulcerative Colitis	Crohn's Disease
Histology	Non-Transmural	Transmural
Endoscopy	Colon	Mouth - Anus
Genetic	IL10, ECM1, etc	ATG16L1, NOD2, etc
Adaptive Immune	T-helper 2	T-helper 1

THE METAGENOMICS DIAGNOSIS FOR IBD CHALLENGE (MEDIC)

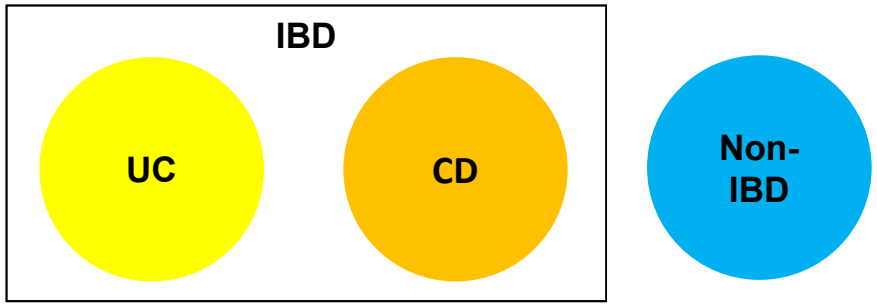
Aim – MEDIC

The challenge aims to **investigate the diagnostic potential of metagenomics data**

1) **to classify IBD patients and non-IBD subjects**
2) within the IBD category, to **attempt to classify subjects with ulcerative colitis (UC) and Crohn's disease (CD)**

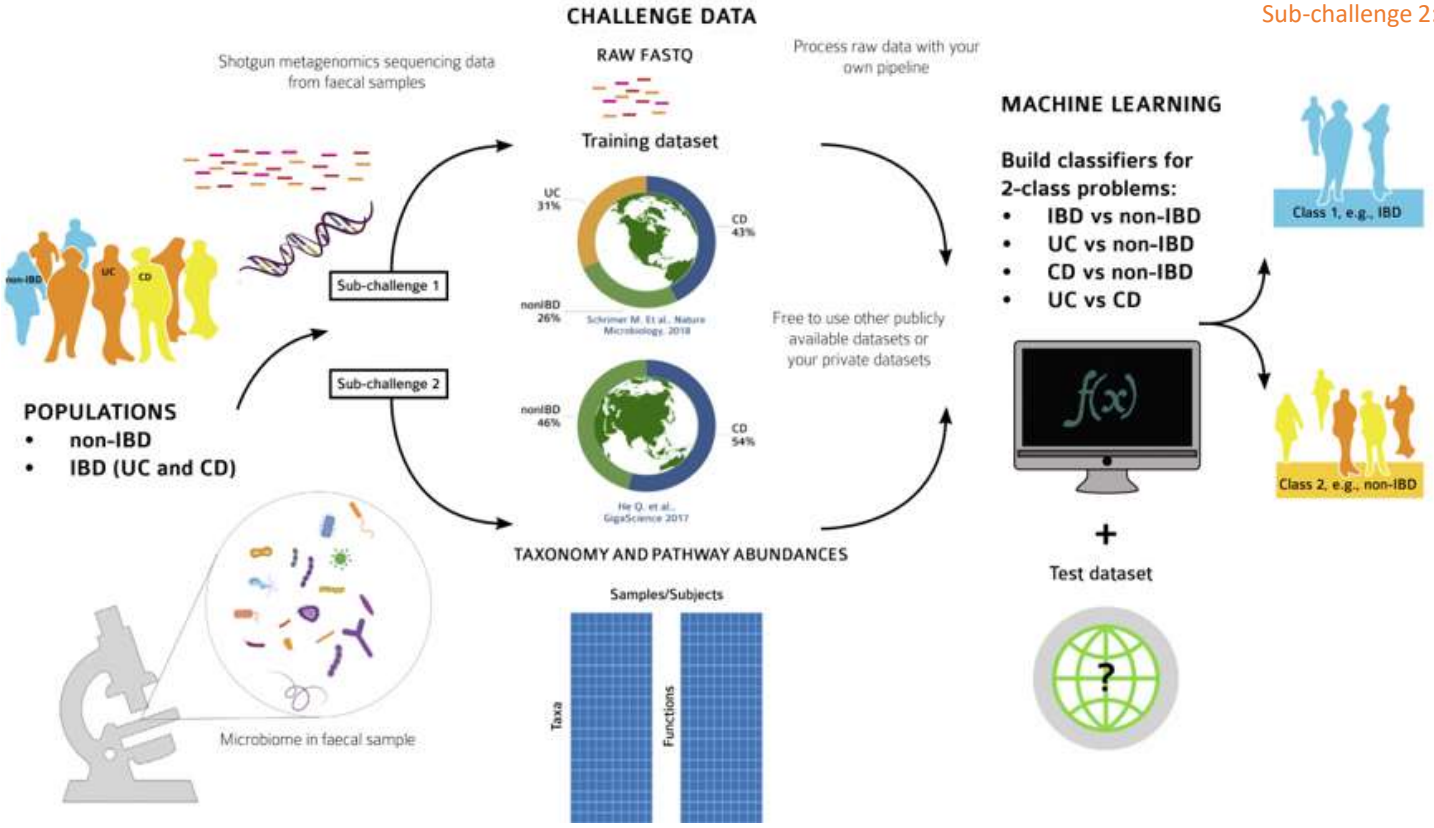
More specifically, the challenge poses four 2-class problems

- IBD vs non-IBD
- UC vs non-IBD
- CD vs non-IBD
- UC vs CD



The challenge

Sub-challenge 1: Requires access to HPC
Sub-challenge 2: Can be completed on a PC



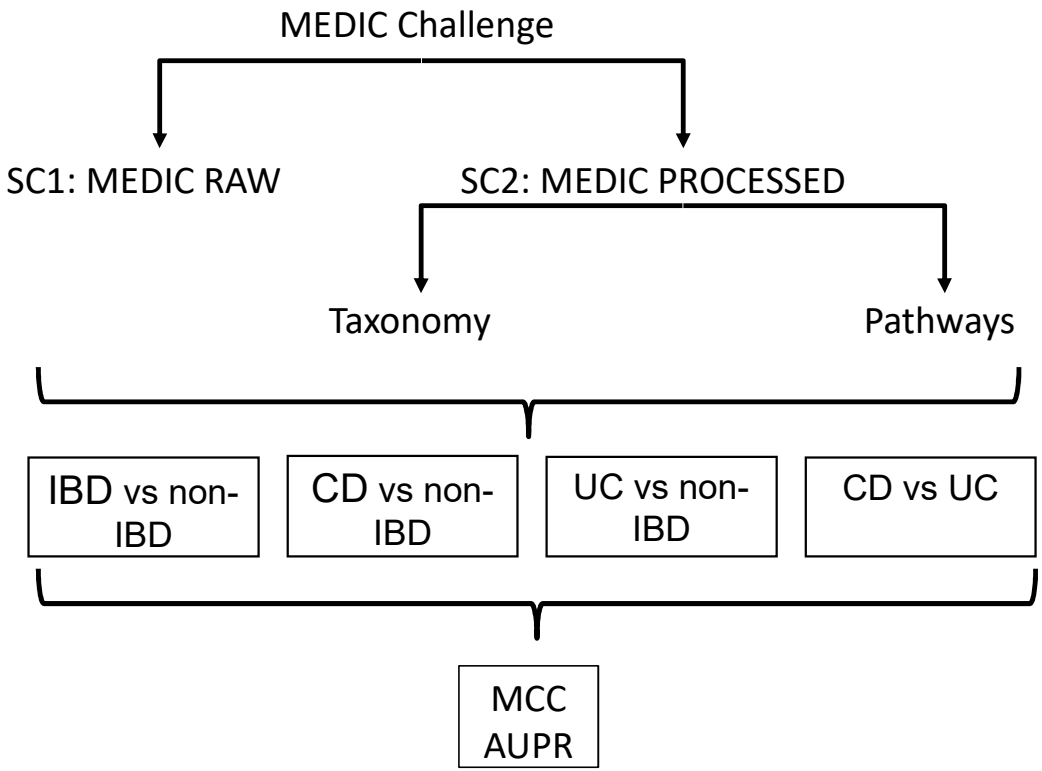
Participants could choose to solve **either one or both sub-challenges.**

SUBMISSIONS AND SCORING

Scoring Procedure

- External and independent scoring review panel (SRP) to approve the scoring strategy before challenge closure
- Metrics and aggregation — Defined upfront and disclosed after challenge closure to avoid development of predictive models optimized for specific metrics
- Anonymized submissions → scorers were blinded to team identity
- After scoring, approval of scoring results and final team ranking by SRP
- Awards for the top 3 best-performing teams for each sub-challenge

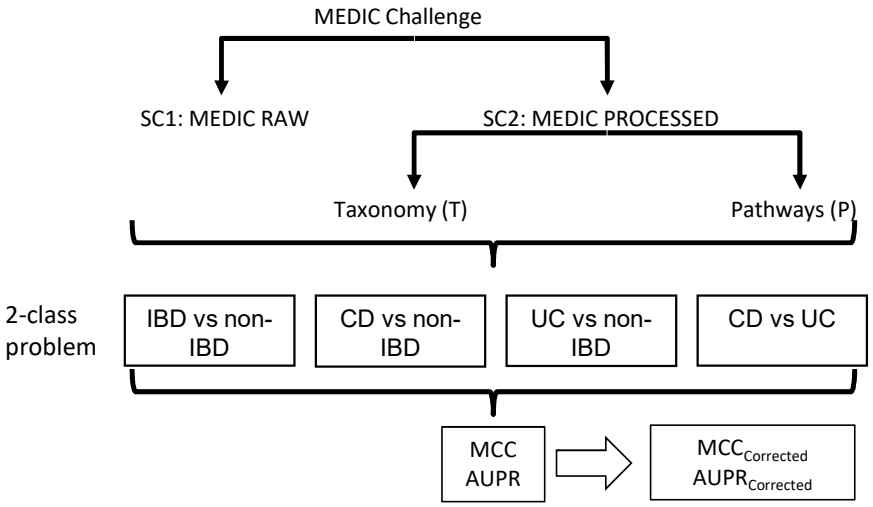
Prediction evaluation (1)



- **2 sub-challenges**
- **2 feature matrices for sub-challenge "MEDIC PROCESSED"**
- **4 two-class problems**
- **2 evaluation metrics**

- Evaluation of prediction randomness
- Score aggregation strategy
- Scoring strategy was developed and approved by the independent scoring review panel before the challenge closed

Prediction evaluation (2)



- For each metric and two-class problem, scores are ranked across teams (the highest score gets the lowest rank)
- For each two-class problem and team, ranks across different metrics will be averaged
- The aggregation of ranks for each team will consist of a weighted sum of ranks giving more weight to the "CD vs UC" two-class problem, which is more challenging
- For each SC, the top 3 teams with the lowest weighted sum of ranks will be declared as the best performing teams after final review and approval by the SRP

$$R_{problem} = \frac{R^{AUPR}_{problem} + R^{MCC}_{problem}}{2}$$

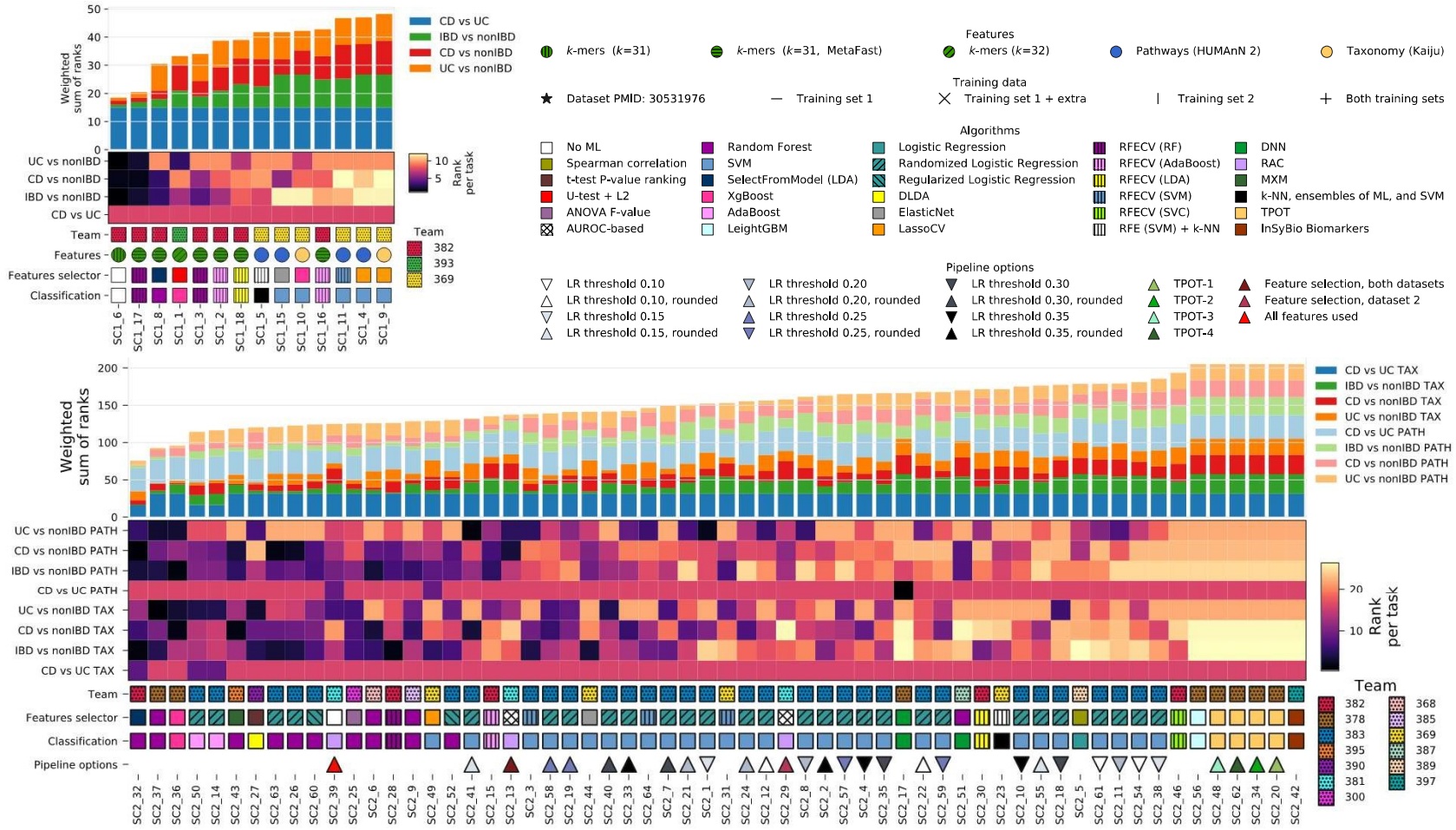
$$Weighted\ Sum\ of\ Ranks_{\ SC1} = R_{IBD\ vs\ non-IBD} + R_{CD\ vs\ non-IBD} + R_{UC\ vs\ non-IBD} + 2 \times R_{CD\ vs\ UC}$$

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

FIRST RESULTS

Submissions summary

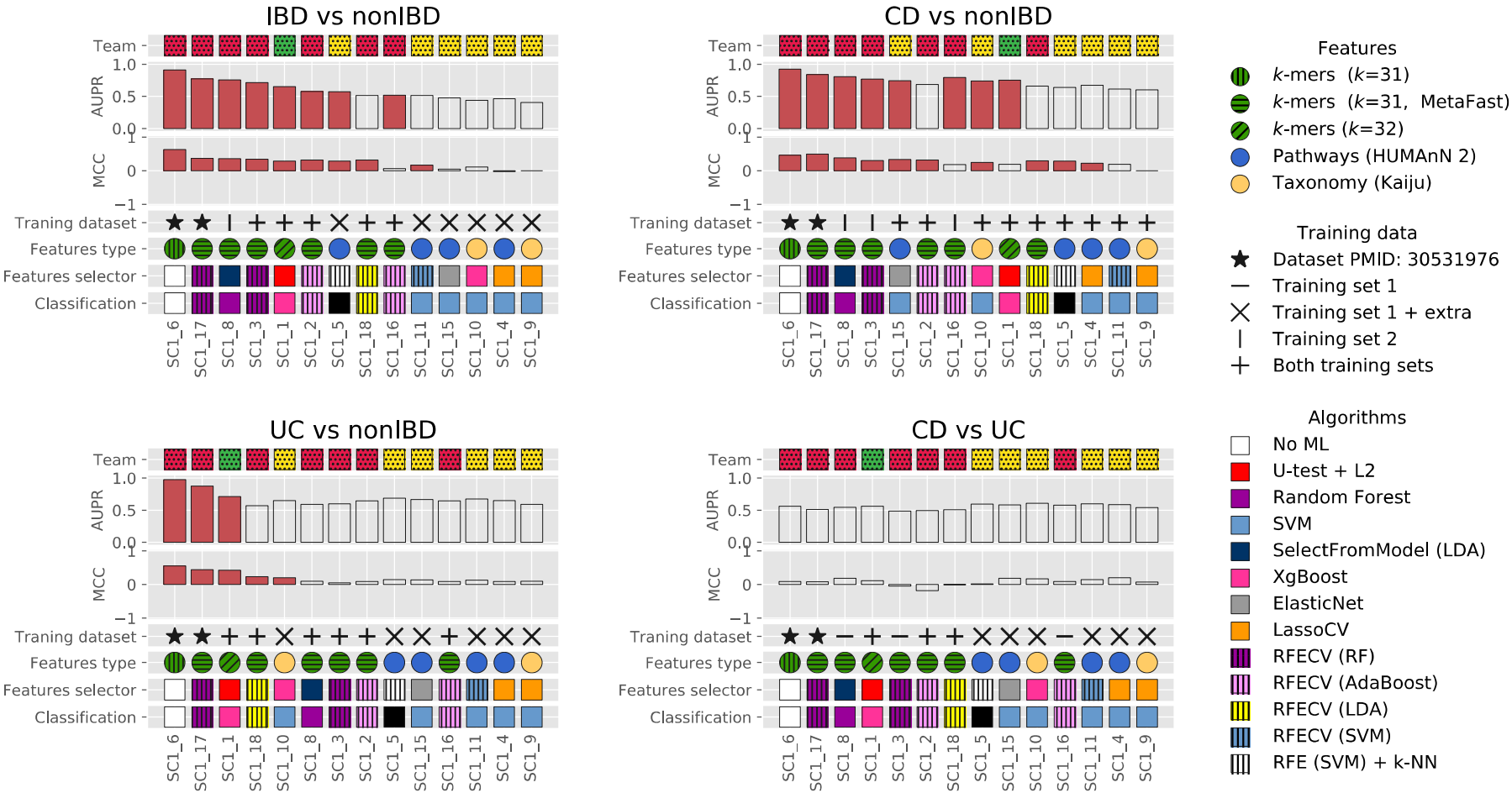
- SC1: **14 submissions** from **3 teams**
- SC2**: 60 submissions** from **13 teams**



CD – Chrohn's Disease
IBD – Inflammatory Bowel Disease
UC – Ulcerative Colitis

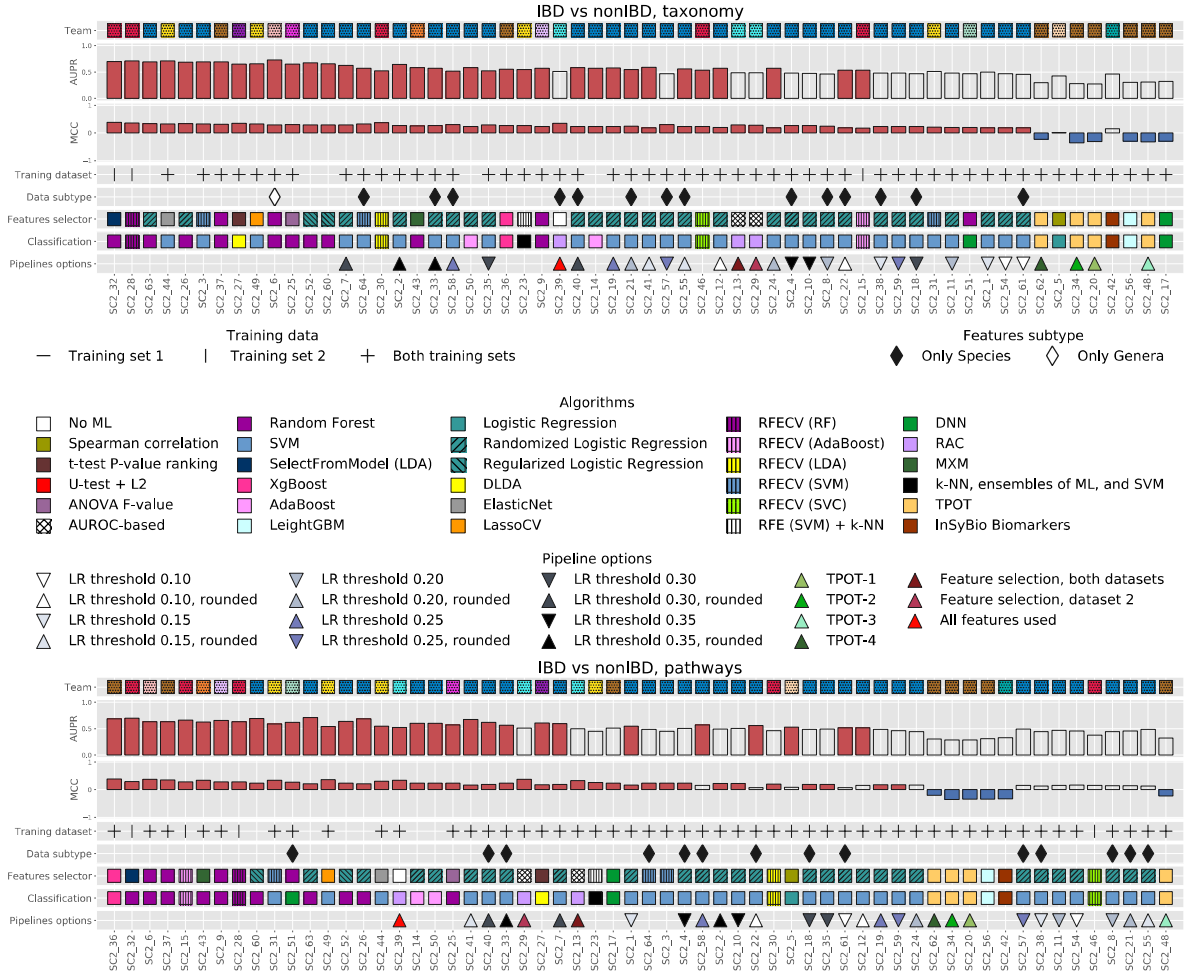
ML – Machine Learning
LDA – Linear Discriminant Analysis
RF – Random Forest
SVM – Support Vector Machine
k-NN – k-Nearest Neighbours
SVC – Support Vector Classifier
DNN – Deep Neural Networks
LR – Logistic Regression

Submissions summary by task (SC1)



Submissions are sorted on the basis of average rank per task.

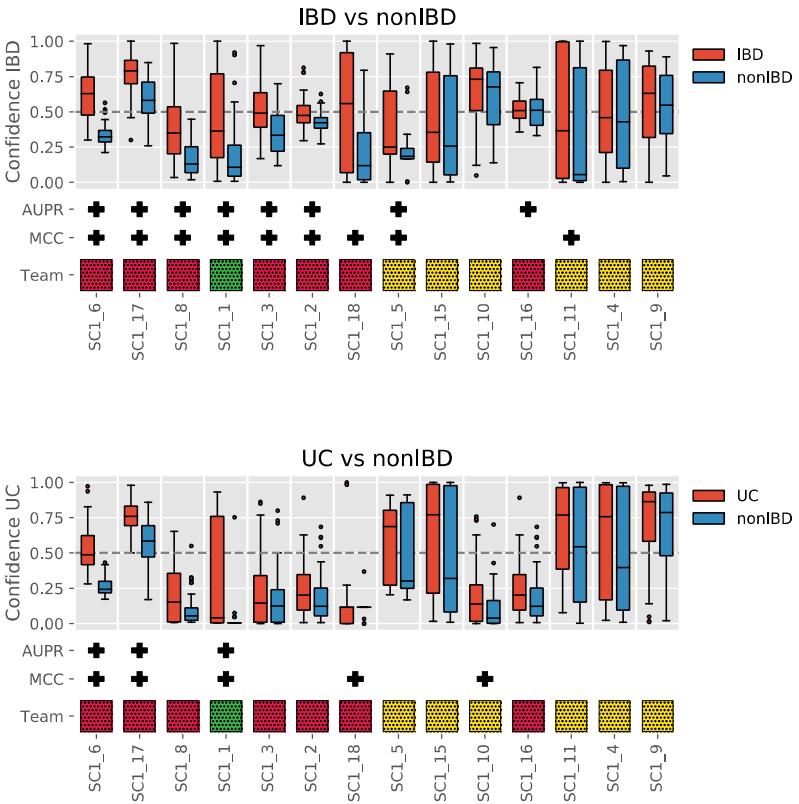
Submissions summary by task (SC2, IBD vs non-IBD task)

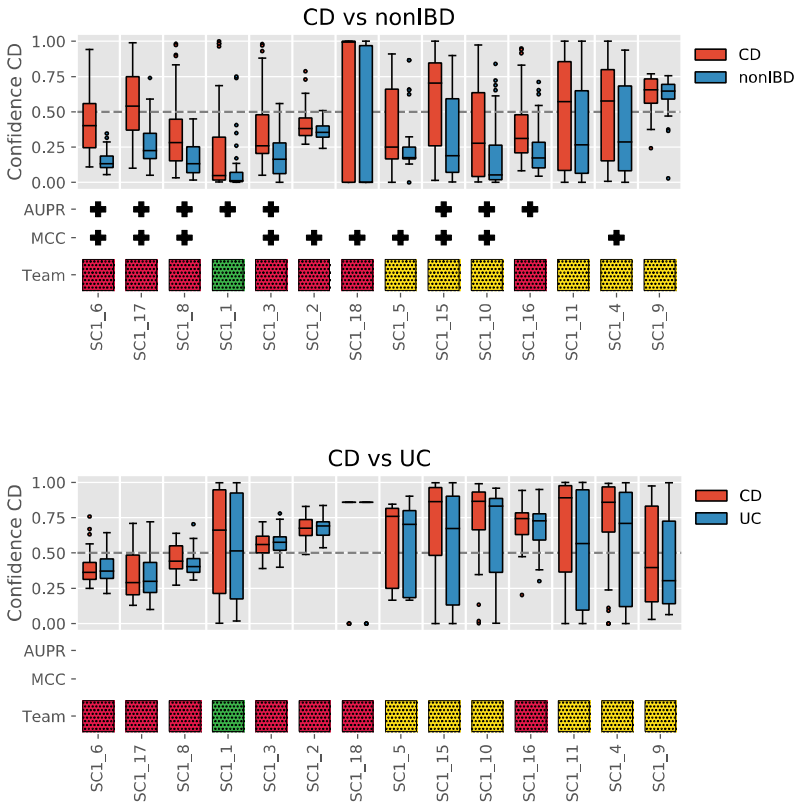


Submissions are sorted on the basis of average rank per task

Although the final ranking was based on overall scoring across data types (taxonomy and pathway) and tasks, the performance of the algorithms varied depending on the task.

Confidence scores (SC1)

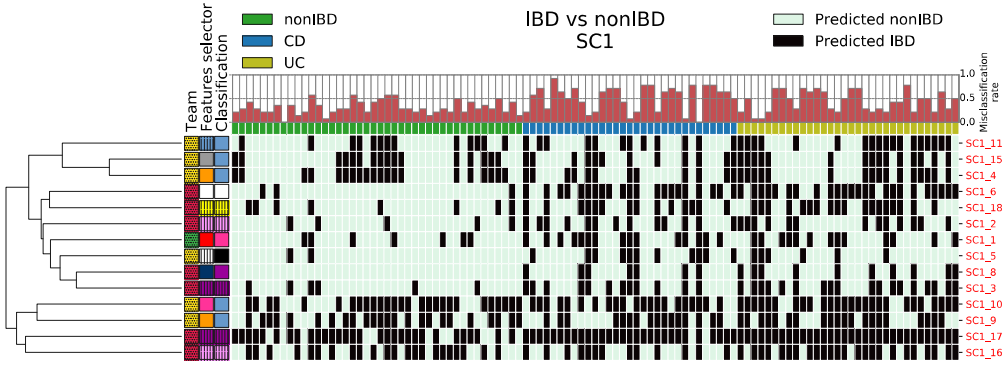


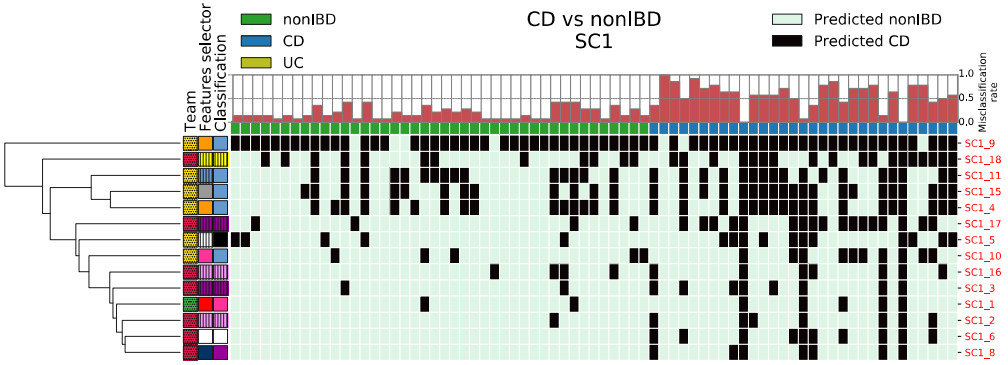


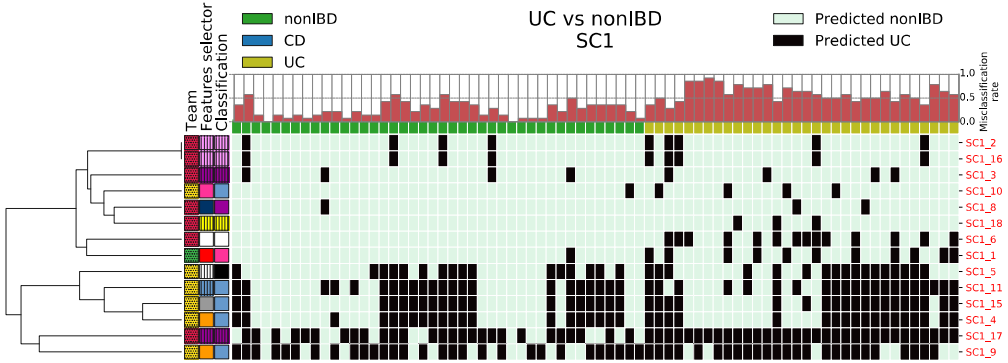
Most classifiers are misclassifying IBD samples.

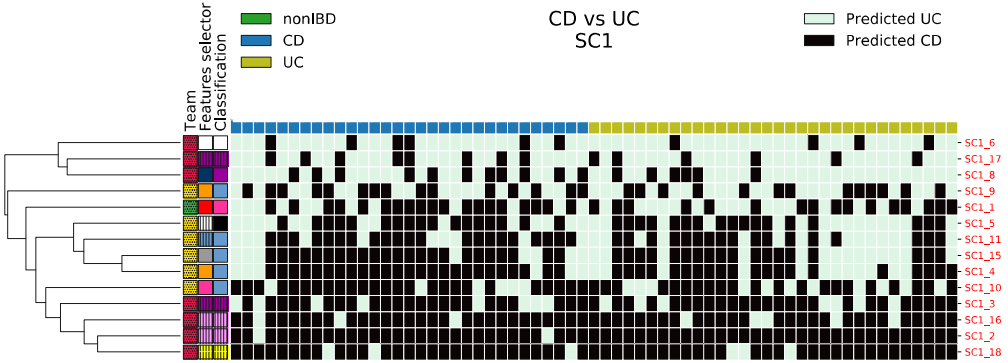
Submissions are sorted on the basis of final performance.

Misclassifications (SC1)





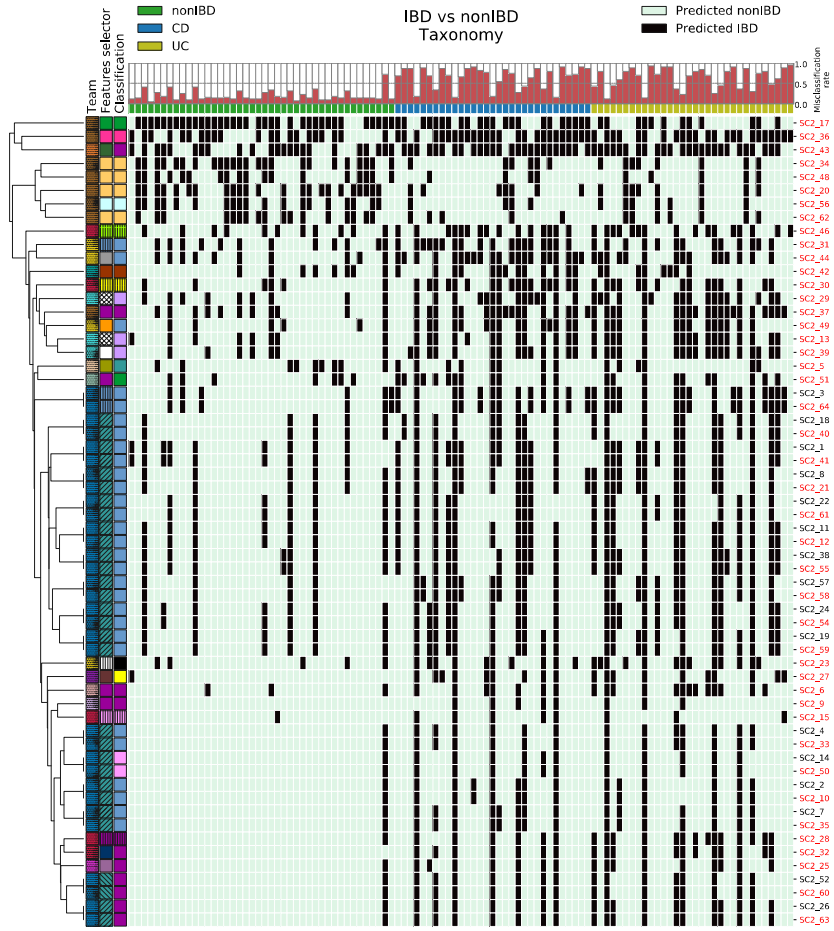


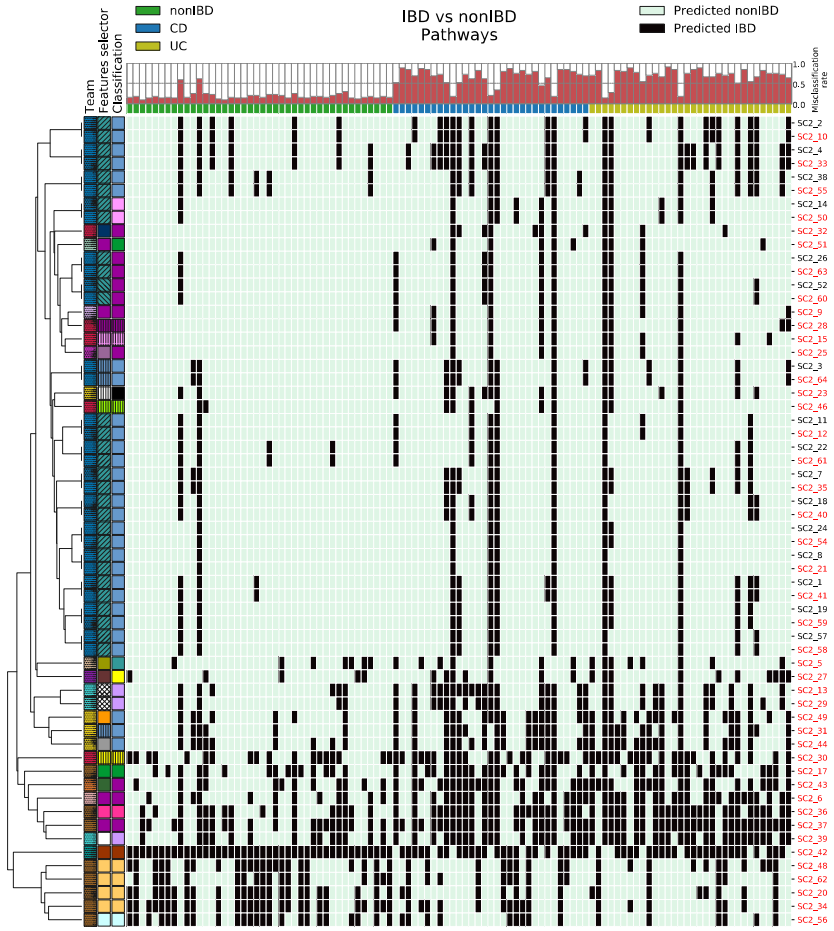


Misclassification patterns are dependent on the algorithm used.

IBD samples were more frequently misclassified than non-IBD samples.

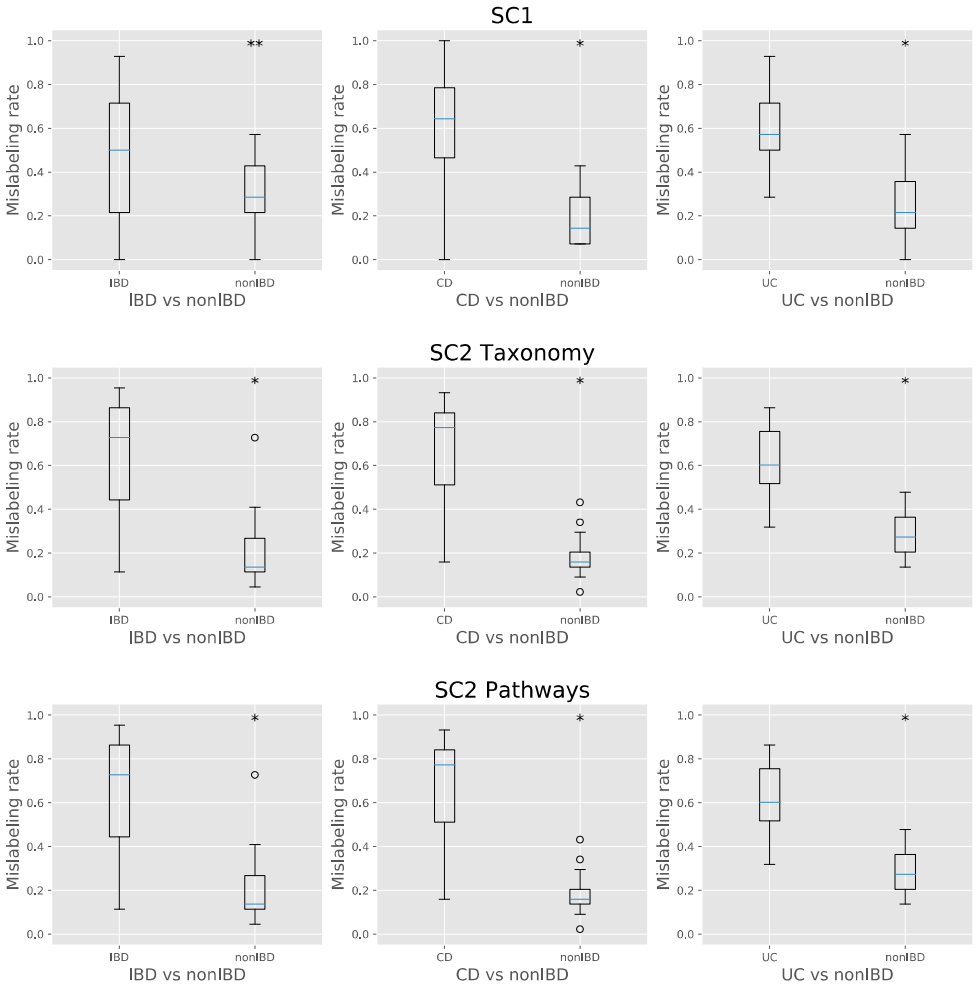
Misclassifications (SC2, IBD vs non-IBD)





Identical binary predictions were identified (shown in black).

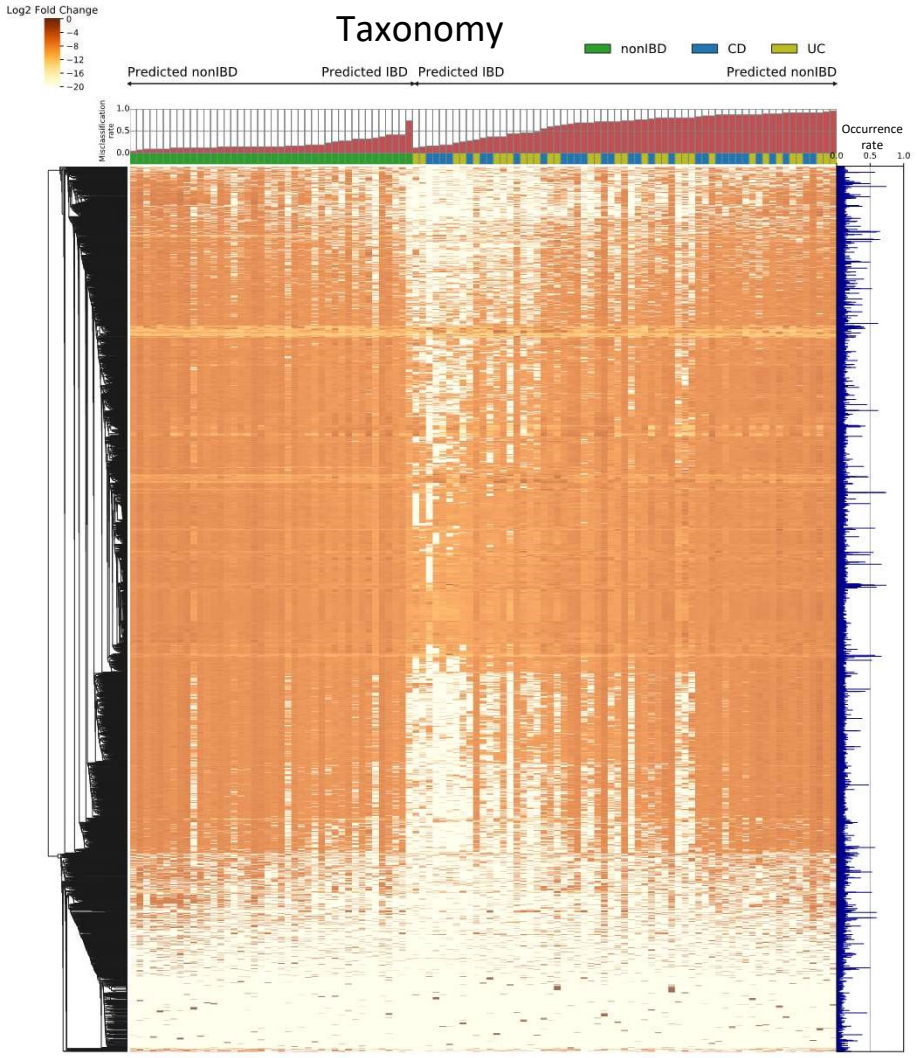
Misclassifications, summary



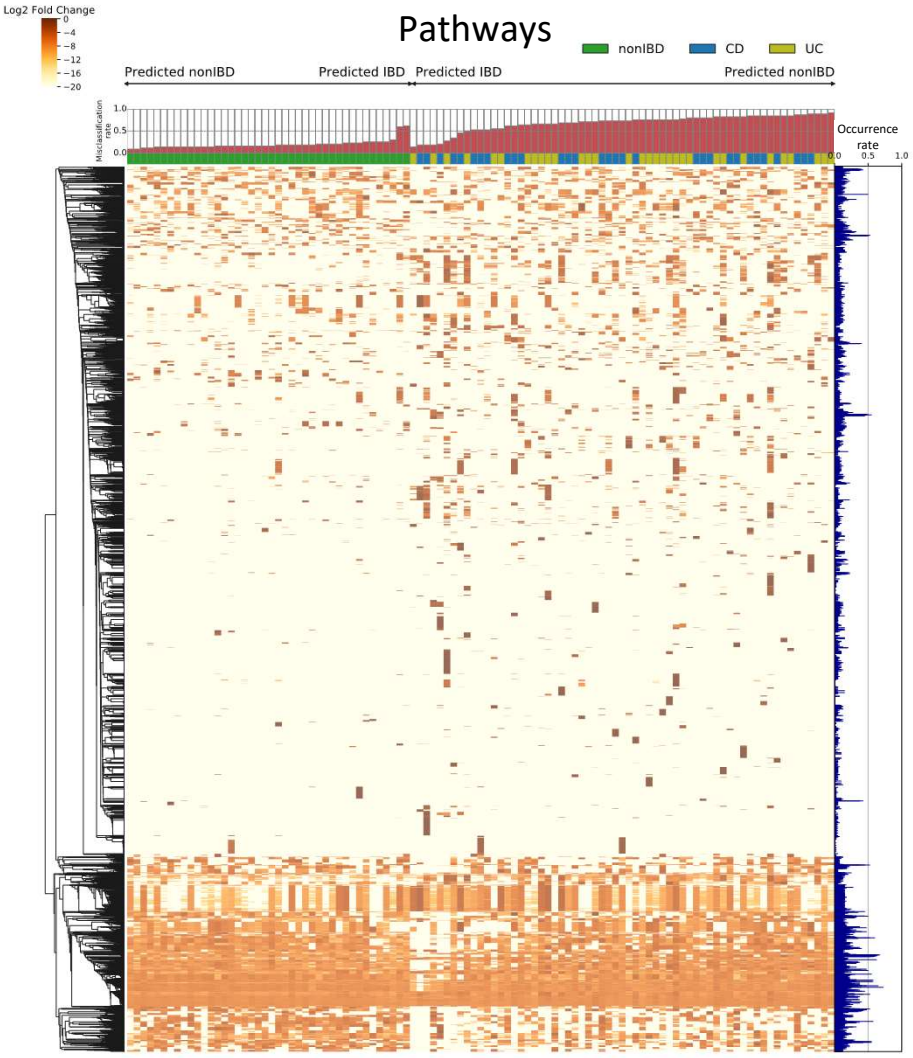
- IBD samples mislabeled statistically more often than non-IBD samples for all data types and tasks.
- The sample misclassification rate and sample clinical metadata were investigated in order to detect the associations; this analysis is still ongoing.

Misclassifications, connection to features (SC2 only)

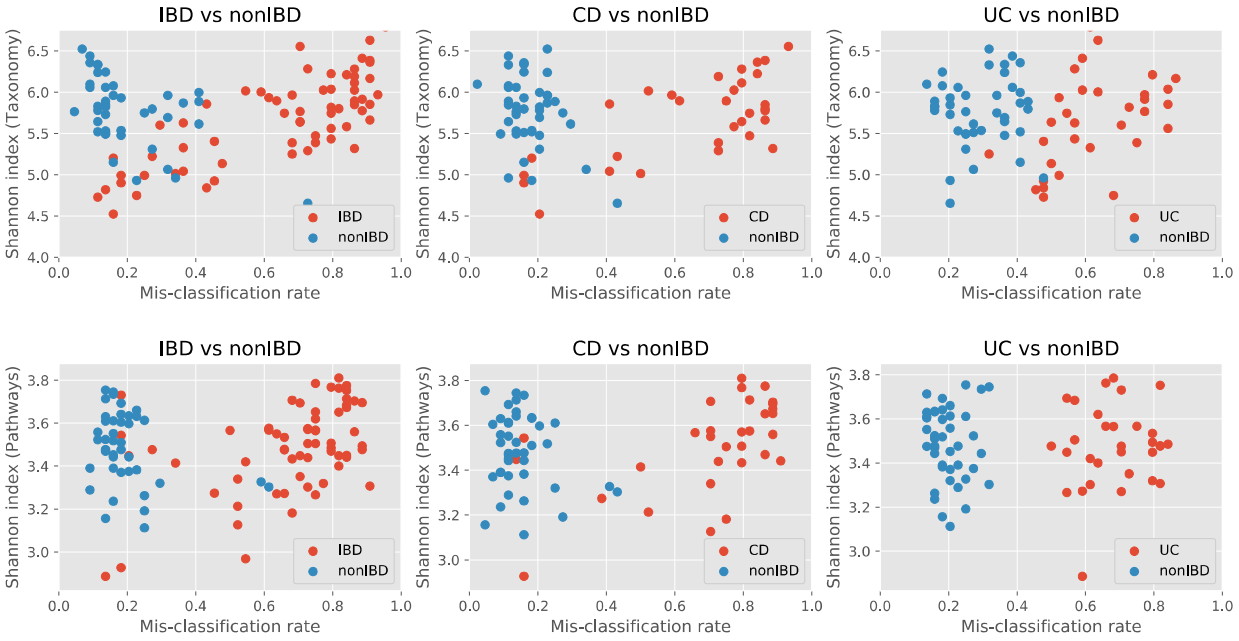
Taxonomy



Pathways



Misclassifications, connection to diversity (SC2 only)



- Positive statistically significant correlation for all tasks and all modes between misclassification rate and diversity (Shannon index) for IBD samples
- Negative correlation (statistically significant only for "IBD vs non-IBD") for all tasks and all data types between misclassification rate and diversity (Shannon index) for non-IBD samples

	Task	Sample type	Correlation coefficient	*P* value
Taxonomy	IBD vs non-IBD	IBD	0.77	**$1.6 * 10^{-13}$**
		non-IBD	-0.52	**0.0004**
	CD vs non-IBD	CD	0.69	**$1.5 * 10^{-5}$**
		non-IBD	-0.36	**0.01**
	UC vs non-IBD	UC	0.44	**0.01**
		non-IBD	-0.05	0.75
Pathways	IBD vs non-IBD	IBD	0.44	**0.0002**
		non-IBD	-0.29	**0.0054**
	CD vs non-IBD	CD	0.53	**0.002**
		non-IBD	-0.21	0.17
	UC vs non-IBD	UC	0.05	**0.048**
		non-IBD	-0.06	0.66

Ensemble of Approaches — Averaging Taxonomy and Pathways Prediction Confidence Values

TAXONOMY	
Sample_001	0.8
Sample_002	0.4
Sample_003	0.7
Sample_004	0.85
.....	
Sample_105	0.2

PATHWAYS	
Sample_001	0.7
Sample_002	0.3
Sample_003	0.1
Sample_004	0.65
.....	
Sample_105	0.1

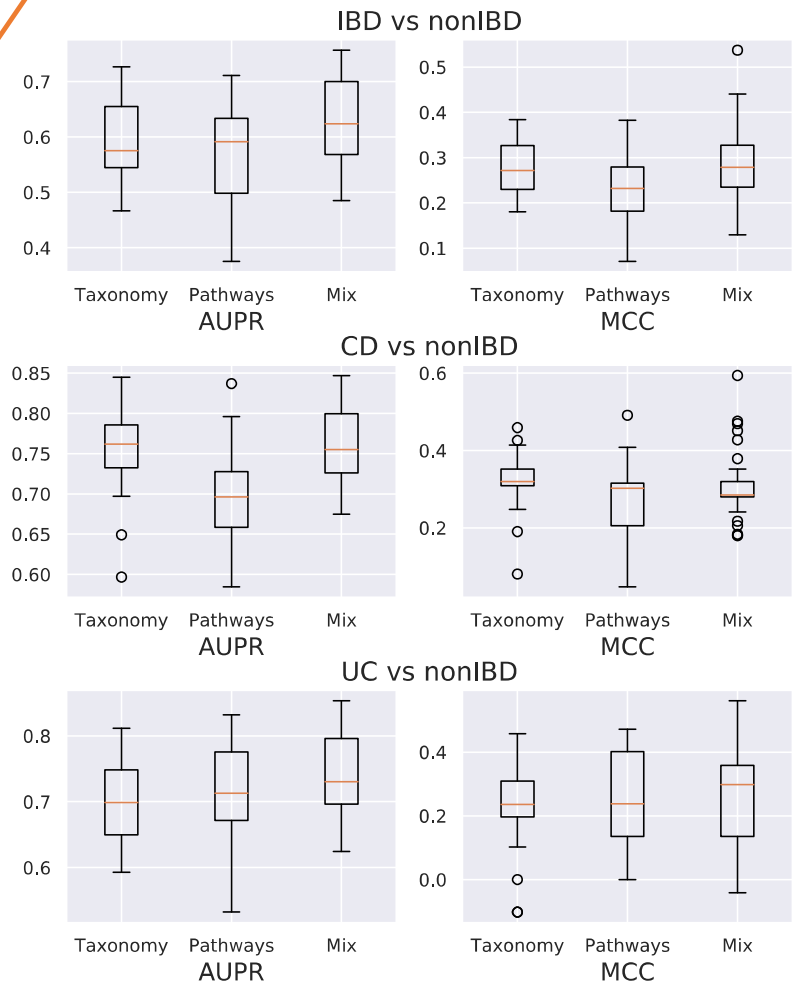
MIX	
Sample_001	0.75
Sample_002	0.35
Sample_003	0.4
Sample_004	0.75
.....	
Sample_105	0.15

Average confidence value per sample

Mixing was done per submission

Only submissions with significant MCC or AUPR for either Taxonomy or Pathways (or both) were considered for this analysis

Confidence values that a sample is non-IBD



Statistical analysis was performed by using the paired-samples Wilcoxon test, with a *P*-value correction for multiple testing.

For all 3 tasks, aggregating taxonomy- and pathway-based prediction confidence values provides a statistically better or similar performance than each separately, suggesting that taxonomy and pathway values are both informative in a complementary way.

CONCLUSIONS

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 samples remains challenging.

- Classification by using *k*-mers-based features showed better performance than classification by using mapping-based features (taxonomy and pathways) provided for SC2

- The type of algorithms that performed best varied depending on the task. 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.

Acknowledgements

Scoring Review Panel

- Dr Laurent Falquet (University of Fribourg, Switzerland)
- Dr Prashantha Karunakar (PES University, Bangalore, India)

sbv IMPROVER Team

- Lusine Khachatryan (Metagenomics Scientist)
- Yang Xiang (Computational Scientist)
- James Battey (Computational Scientist)
- Adrian Stan (Communication)
- Conny Johannsson (Strategy & Communication)
- Gill Den Hartog (Project management)
- Giuseppe Lo Sasso (IBD Scientist)
- Nicolas Sierro (Genomics Lab Manager)
- Carine Poussin (Computational Scientist and Challenge Leader)
- Stéphanie Boué (Project Leader)
- Nikolai V. Ivanov (Technology Facility Manager)
- Julia Hoeng (Program Leader)

THANK YOU FOR YOUR ATTENTION