Crowdsourced-benchmarking of computational pipelines for metagenomic taxonomy profiling – the sbv IMPROVER Microbiomics Challenge

Carine Poussin[1], Nicolas Sierro[1], Vincenzo Belcastro[1], James Battey[1], Stéphanie Boué[1], Fernando Meyer[2], Elena Scotti[1], Alexandros Petrelis[1], Manuel C Peitsch, Nikolai V Ivanov[1], Julia Hoeng[1]

[1]*Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland (part of Philip Morris International group of companies)*
[2]*CAMI, Department of Computational Biology of Infection Research, Helmholtz Centre for Infection Research (HZI), Braunschweig, Germany*

The Microbiota Composition Prediction Challenge

BACKGROUND

A growing body of evidence suggests that the equilibrium of microbiome communities and their interaction with their host plays an important role in maintaining host health. **Changes in microbiome communities (dysbiosis)** may be a cause or a consequence of the development of many diseases.

The choice of adequate computational metagenomics methods is **decisive in determining microbial taxonomy profiles and functions accurately**. However, there is no consensus yet about the best analytical and computational approaches to use.

Crowdsourcing initiatives such as Critical Assessment of Metagenome Interpretation (CAMI) (http://www.cami-challenge.org/) have started to benchmark algorithms focusing on specific aspects of metagenomics data analysis (1).

The sbv IMPROVER crowdsourcing project (sbvimprover.com), as a means to verify methods and data in systems biology, has already shown its usefulness in benchmarking computational methods through crowdsourcing.

THE CHALLENGE

To build and expand upon what has been done by CAMI, the design of the sbv IMPROVER Microbiomics Challenge focuses on the influence of sample complexity (number of species) and sequence bias (AT vs. GC-rich) on the quantification of microbial communities at various taxonomic levels based on shotgun sequencing data.

GOAL

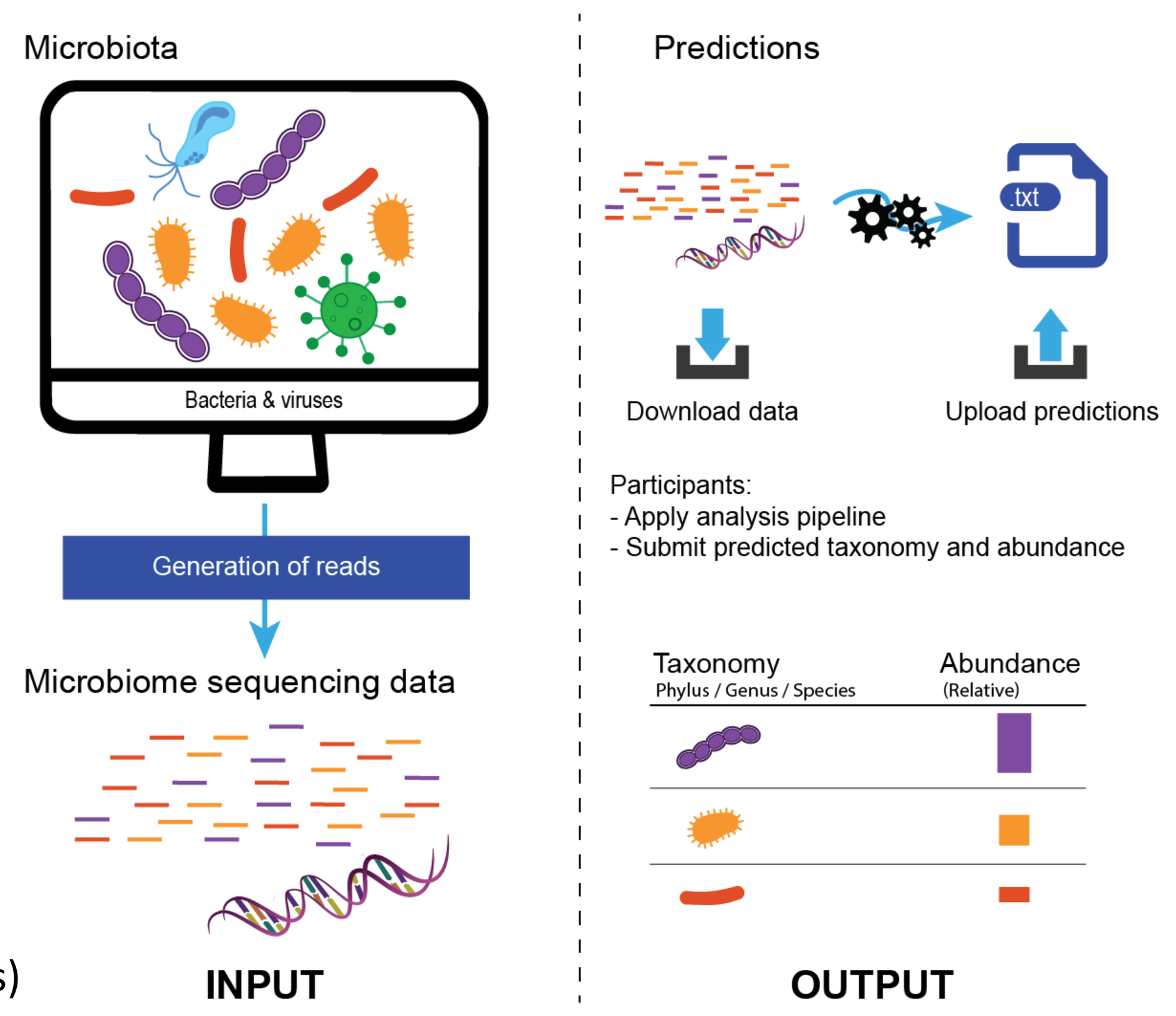
The challenge aims to assess the performance of **metagenomics computational analysis pipeline(s) as a whole to recover relative abundance and taxonomy assignment of bacterial communities**.

Scientific questions
- Which pipelines best recover bacterial community composition and relative sequence read abundance at phylum, genus and species taxa rank?
- Do technical biases and specific microbial composition affect the performance?

Timelines
The challenge was open from mid-November 2017 to mid-June 2018 (7 months)

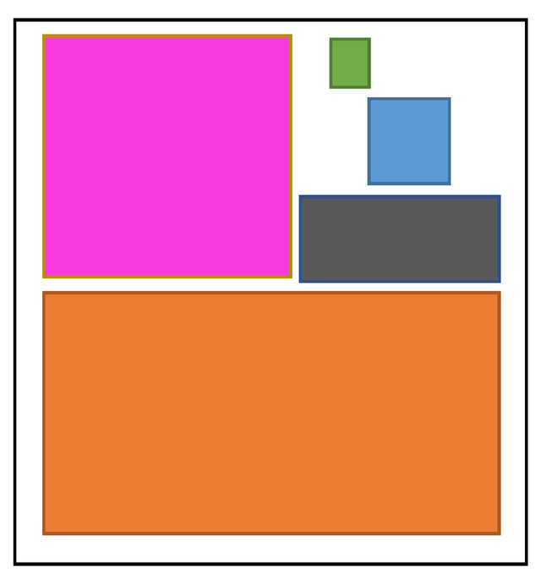
Rules
Eligibility for scoring depended on:
- Submission completeness
- Compliance to data format and challenge rules

Best performers' award
Travel bursary of $2'000 for a conference (3 best teams)



DATASET

Nineteen paired-end reads samples were provided to the participants (two fastq files per sample). Among these 19 samples, 15 samples have been generated computationally using the ART (2) set of simulation tools using parameter settings to simulate next-generation sequencing reads from an Illumina HiSeq4000 sequencer ("Simulated"), and the four remaining samples correspond to actual Illumina HiSeq4000 sequencing data from a commercial standard DNA control with known bacteria mix ("Real").



Sequencing Reads	# Species	# Samples	Characteristic
Simulated	10 (Complexity 1)	3	unbiased
	40	3	GC-rich bacteria (biased)
	120	3	AT-rich bacteria (biased)
	500	3	
	1000 (Complexity 5)	3	+ Mouse reads
Real	10*	2 (lib. 1)	DNA control (ZymoBIOMICS) – 2 library preparations
	10*	2 (lib. 2)	

(*8 bacteria species, 2 yeast species) =19 samples in total

SCORING

Anonymized participants' submissions were split by samples and taxonomic ranks, and evaluated against the corresponding gold standards, i.e., known composition of the respective samples using metrics implemented by CAMI in the OPAL (https://github.com/CAMI-challenge/OPAL) software as defined below. Various reference computational pipelines were evaluated by us and CAMI in addition to participants' submissions. These latest predictions, not eligible for the challenge awards, were nevertheless scored for comparison with other approaches tested by challenge participants.

Qualitative/Binary classification metrics: allow to assess **how well a particular method detects the presence or absence of an organism** in comparison to the gold standard.
- **F1 score** [0-1]: is the harmonic mean of precision and recall. As it does not put weight on true negatives, which are predominant in microbiome datasets, it is suitable in this context.

Quantitative/Abundance metrics: allow to assess **how well a particular method reconstructs the relative abundances** in comparison to the Gold Standard.
- **L1-norm**: distance shows the distance between relative abundance vectors.
- **Weighted UniFrac**: distance metric for comparing the composition of microbial communities, which incorporates phylogenetic information (3).

Score aggregation for final ranking: consisted of weighted sum of ranks of each sample-taxa-metric rank per team. The best performer was the team with the lowest aggregated rank.

RESULTS

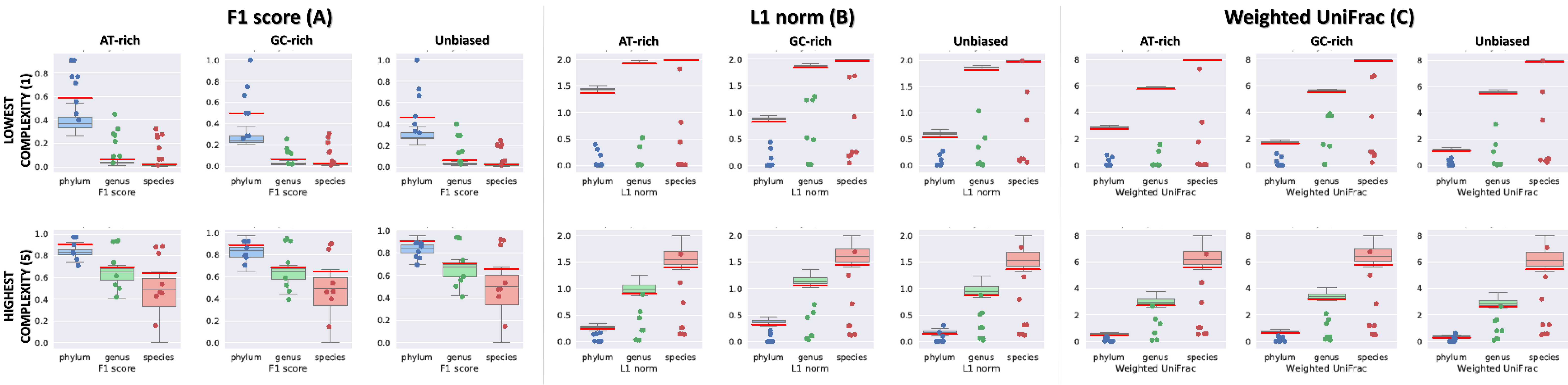
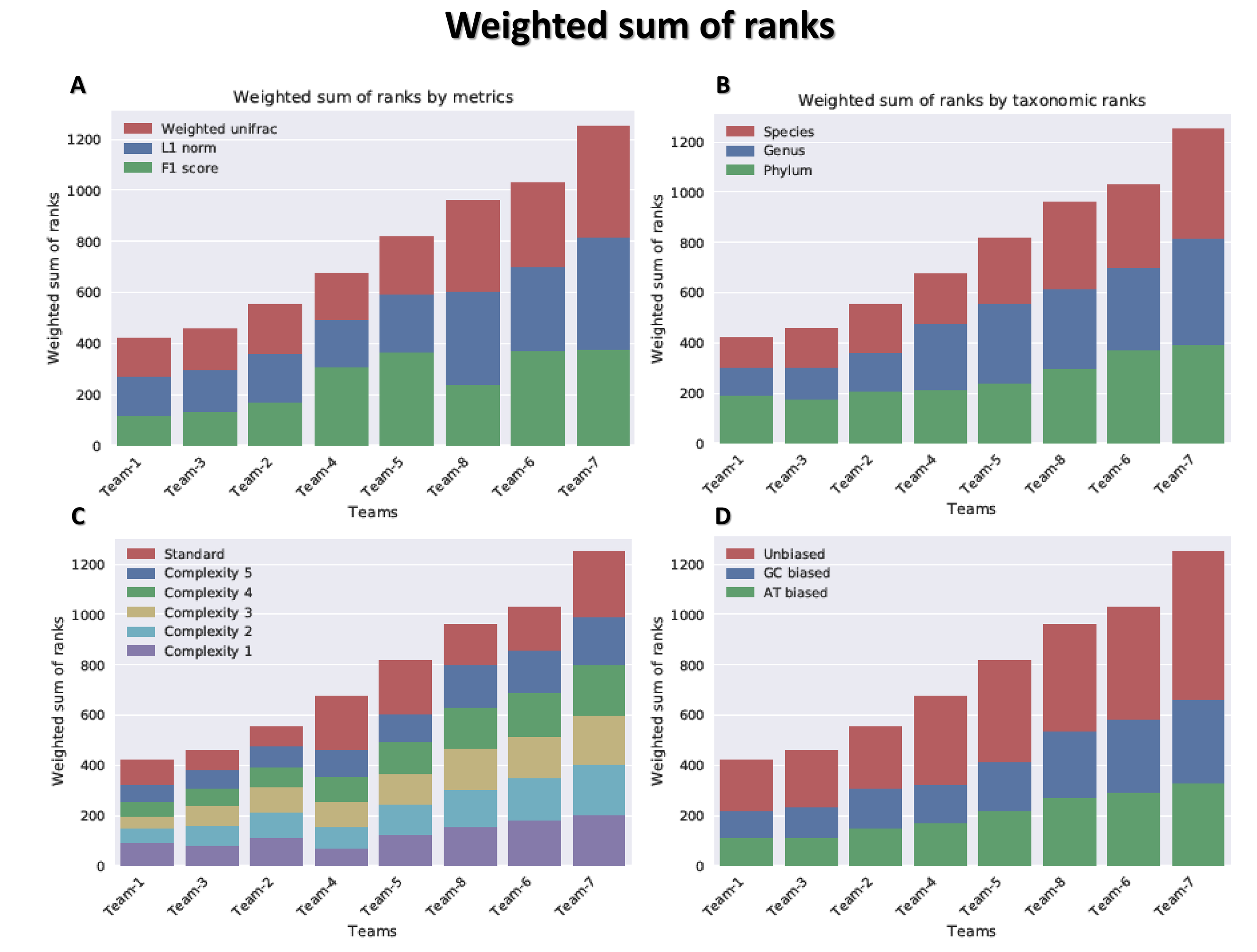


Figure 1. F1 (the higher the better), L1-norm (the lower, the better) and weighted UniFrac (the lower, the better) scores for predictions from challenge participants. F1 (A), L1-norm (B) and weighted UniFrac (C) scores were compared to the score distribution calculated from 10'000 random predictions. Results for the lowest (1) to the highest complexity (5) for the simulated biased (AT- and GC-rich) and unbiased datasets are shown here. The boxplots on each plot represent the distribution of scores calculated from the 10'000 random predictions. The score corresponding to the 95th percentile of the random prediction score distribution is displayed as a red line for each taxonomic level. Teams' predictions below or above the red line depending on the metric are considered as random predictions.

Weighted sum of ranks



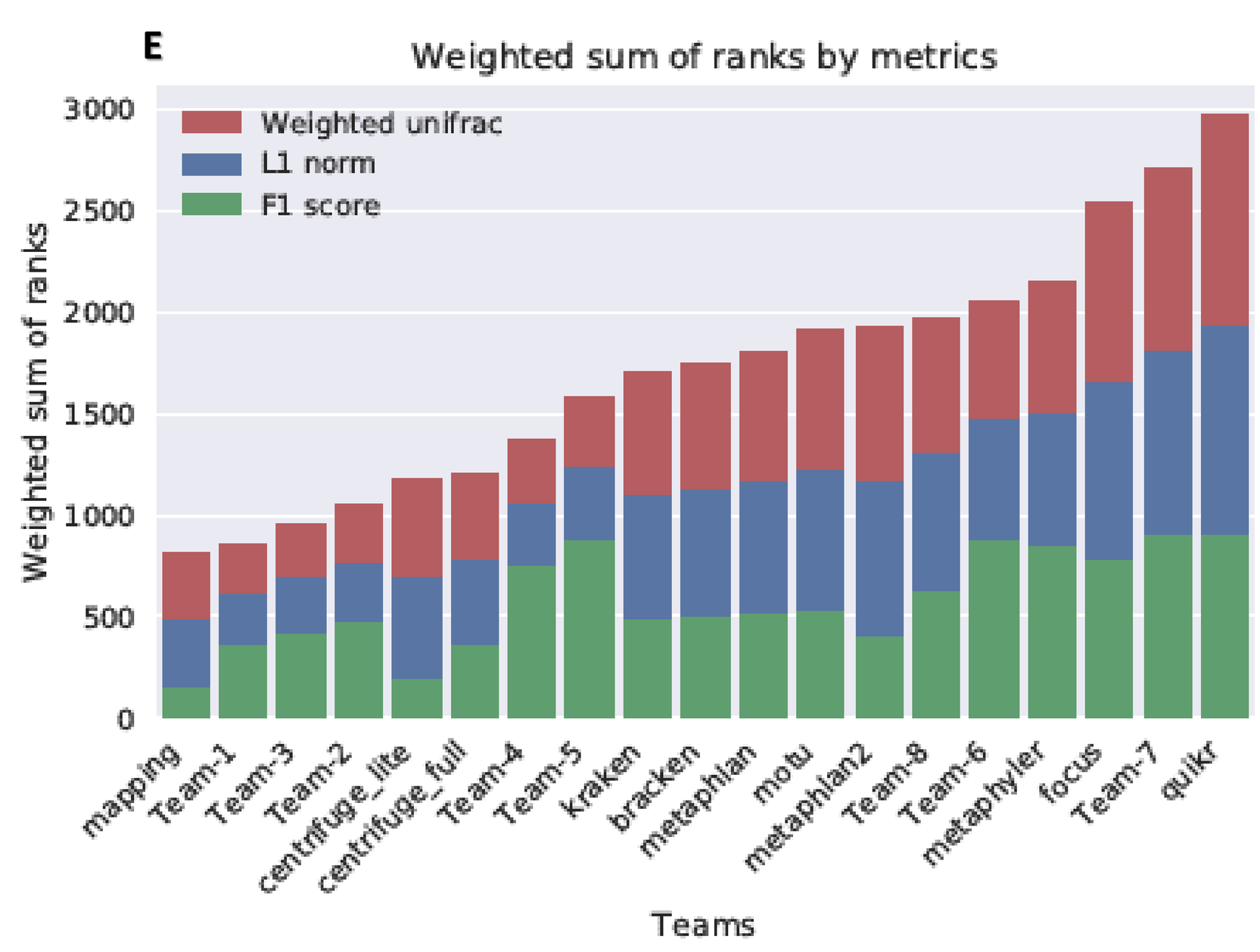
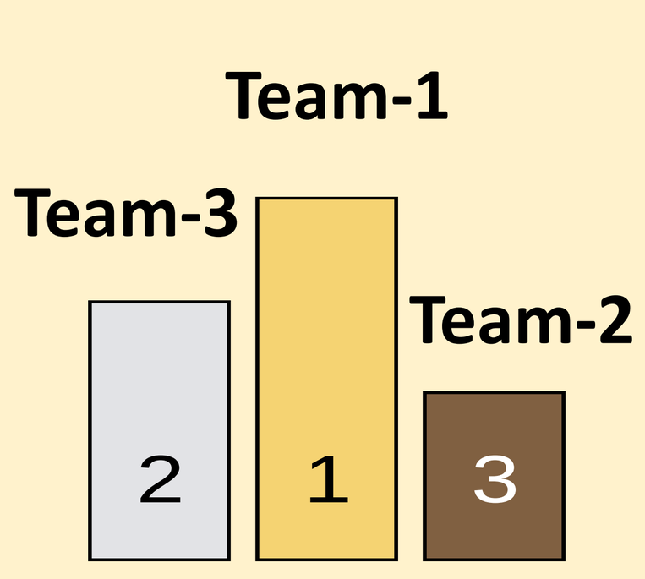


Figure 2. Weighted sum of ranks (WSR) showing the contribution by metrics (A), taxonomic ranks (B), complexity (C) and bias status (D). The teams are ordered on the barplot from the lowest to the highest sum of ranks (Real data = Standard; Simulated data = complexity 1-5). (E) WSR plot including computational pipelines used by participants, and the ones tested by CAMI and PMI on the dataset.

Best performers



Summary

- Eight submissions from Japan, India and Armenia were scored
- A majority of predictions were better than random when assessing relative abundances, while more predictions appeared to be random when only looking at presence or absence of a taxa
- The percentage of teams'predictions evaluated not better than random (171 scores computed per submission) ranged from 0.0058% to 0.43%
- The teams used various computational approaches
- The most performant pipeline was based on the Braken software
- For teams using the same softwares, differences in performances arrose from the use of different genome databases, data preprocessing (e.g. filtering) and/or software parameters
- Post-challenge analysis will focus on understanding these differences and provide new learnings

References & Acknowledgements

1. Sczyrba et al., Nature Methods 14, 2017
2. Huang W et al., Bioinformatics 28(4), 2012
3. Lozupone et al., Isme J 5(2), 2011

www.sbvimprover.com
Questions? Contact Us
sbvimprover.RD@pmi.com

- We thank Prof. Alice McHardy (Helmholtz Centre for Infection Research, Germany) and Dr Luiza Cutillo (Department of management and quantitative Studies - University Parthenope of Naples, Italy) part of the sbv IMPROVER Scoring Review Panel for the Microbiomics challenge.
- We also thank Laure Cannesson and David Page for the project management and communication, and Sylvain Gubian for programming support at PMI.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Competing Financial Interest
The research described in this poster was funded by Philip Morris International