Microbiome analysis – a review of current recommendations and utilization of a proven crowdsourcing solution to benchmark methods.

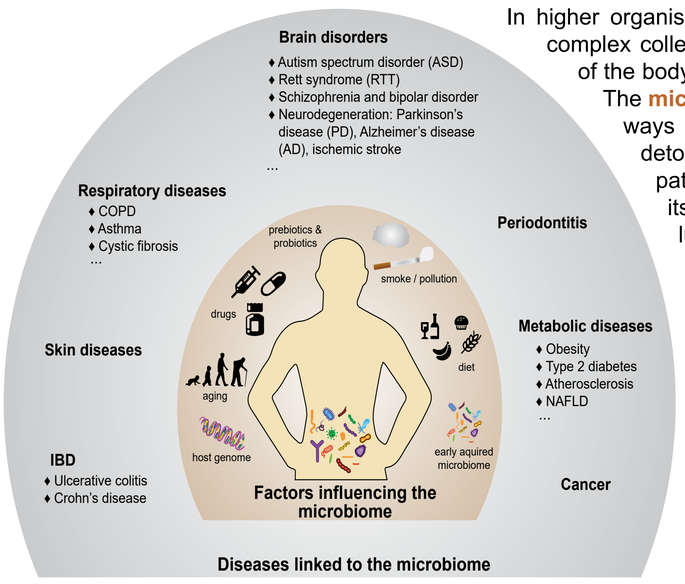
James Battey, Stéphanie Boué, Carine Poussin, Nicolas Sierro, Elena Scotti, Vincenzo Belcastro, Alexandros Petrelis, Giuseppe Lo Sasso, Raffaella Maria Gadaleta, Nikolai V. Ivanov, Julia Hoeng
PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland (Part of Philip Morris International group of companies)

Dysbiosis, a disturbance in gut microbial equilibrium, provokes dysregulation of the adaptive immune response in the gut and is recognized as one of the main contributing factors in the development of Inflammatory Bowel Diseases (IBD), including Ulcerative Colitis and Crohn's Disease. Metagenomics studies in IBD patients and control subjects contribute to unravelling the involvement of intestinal micro-organisms in IBD pathogenesis. Key aspects of microbiomics study design and conduct as well as data analysis will be illustrated. Importantly, there is no consensus about the best analytical and computational approaches to use.

The sbv IMPROVER crowdsourcing project, developed by Philip Morris International as a mean to verify methods and data in systems biology, has already proven its usefulness in benchmarking computational methods used, for example, in diagnostic signature discovery or the assessment of species translatability. The design of the latest sbv IMPROVER challenge focuses on the influence of sample complexity and/or sequence bias on the quantification of microbial communities at various taxonomic ranks based on shotgun sequencing data. In anticipation of the results from the challenge, preliminary results obtained with a few methods implemented by the challenge organizers will be shared.

The sbv IMPROVER challenge will contribute to learning more about specific aspects of data analysis in microbiomics, which may prove key in developing new routes for diagnosis and therapy in a number of disease areas.

Importance of the Microbiome in Health and Disease



From Scotti et al. Exploring the microbiome in health and disease: Implications for toxicology. *Toxicology Research and Application* (2017)

In higher organisms, such as humans, the microbiome comprises a complex collection of microorganisms that colonize different parts of the body including the gut, mouth, genitals, skin and airways. The **microbiome interacts with its host** in several important ways (e.g., assisting in the bioconversion of nutrients and detoxification, supporting immunity, and protecting against pathogenic microbes). It is now recognized that through its close interaction with the nervous system and the lungs, the **microbiome has a strong influence on general health**. The function of the indigenous microbiota can be influenced by many factors, including genetics, diet, age, and toxins. **Dysbiosis,** the disruption of the delicate balance between the microbiome and its host, has been associated with increased risk for a number of diseases, including disorders associated with obesity, chronic inflammatory diseases, cardiovascular diseases, cancers, and neurodegenerative disorders (Scotti 2017).

Advances in genome sequencing technologies have enabled progress in the characterization of microbial diversity, leading to a rapid expansion of the field known as **metagenomics**: the study of DNA of a microbial community.

In the future, microbial abundances could be used as markers for **disease diagnostics**. In addition, the understanding of the importance of the microbiome for human health has led to the emergence of **novel therapeutic approaches** for a variety of conditions. These are focused on the manipulation of microbiota, either by restoring beneficial microbes that are missing or by reducing or eliminating those that are associated with pathology. Many of these therapeutic possibilities are in the early stages of development and the potential of microbiomics for the diagnosis and treatment of human disease is just beginning to be realized.

sbv IMPROVER

Based on the principles of **crowdsourcing** and **collaborative competition**, the sbv IMPROVER project is designed as a series of open **scientific challenges**, in which computational methods are benchmarked and conclusions related to scientific problems of interest in the systems biology and/or toxicology fields are scrutinized rigorously (Meyer 2011).

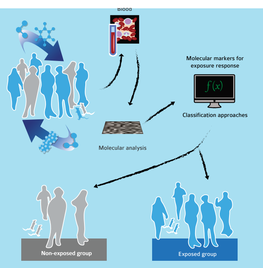
In strategically engaging the crowd, sbv IMPROVER **facilitates enhanced dialogue within the scientific community, transparency of research processes, and open innovation in scientific discovery**. The project advances the credibility of scientific techniques and complements the classical peer review process with a rigorous benchmarking of computational methods and assessment of conclusions. Computational challenges leverage the wisdom of the crowd allow to benchmark methods for specific tasks, such as signature extraction or sample classification. Four challenges have already been conducted successfully, and it has been confirmed that the agglomeration of predictions often leads to better results than individual predictions and that methods perform best in specific contexts.

A challenge is posed by defining clear objectives and rules and providing data to the participants. The part of the data that needs to be predicted, called the "Gold Standard" (true values), is kept hidden from the participants and is used in combination with pre-defined scoring metrics to assess the performance of anonymized participants' prediction submissions. Scoring results and team rankings are submitted to an external and independent Scoring Review Panel of experts for review and final approval. The results, conclusions and lessons learned from the challenge are shared with participants and with the scientific community through conference and symposium presentations and in peer-reviewed publications.

To learn more about the project and associated publications, please visit www.sbvimprover.com.

Systems Toxicology Challenge

The Systems Toxicology Challenge aimed to verify that robust and sparse human-specific and species-independent gene signatures of exposure response can be extracted in whole blood gene expression data from humans and rodents to predict exposed and non-exposed group labels (Poussin 2017).



Special session at ISMB2016 and Symposium 2016, Orlando (USA)

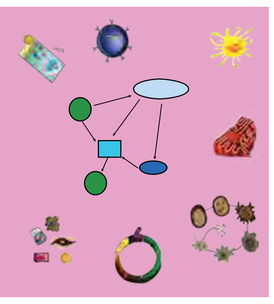
Species Translation Challenge

The Species Translation Challenge aimed to verify that changes in phosphorylation status and gene set activation induced by cellular response to 52 different perturbations in human cells can be predicted to a certain extent, given responses generated in rat cells (Poussin 2014).

Symposium 2013, Athens (GR)

Network Verification Challenge

The Network Verification Challenge aimed to verify biological network models to ensure their relevance to lung biology and chronic obstructive pulmonary disease (sbv IMPROVER project team 2015).

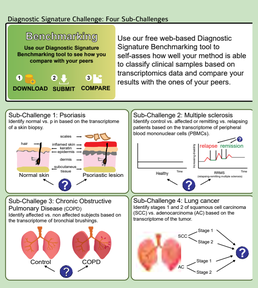


Symposium 2014, Montreux (CH)
Symposium 2015, Barcelona (ES)

Diagnostic Signature Challenge

The Diagnostic Signature Challenge aimed to assess and verify computational approaches that classify clinical samples based on transcriptomics data.

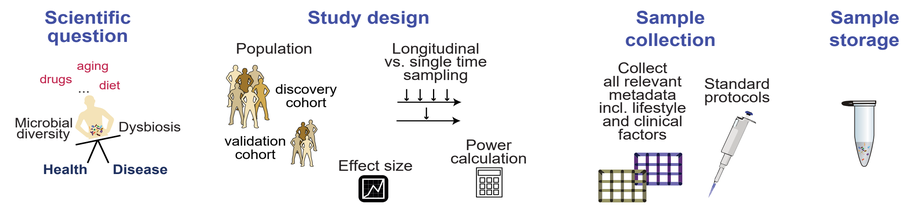
The high quality of predictions strongly confirmed the approach values (Tarca 2013).



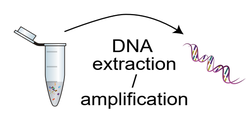
Symposium 2012, Boston (USA)

Microbiome Analysis

1. Study design & sample collection



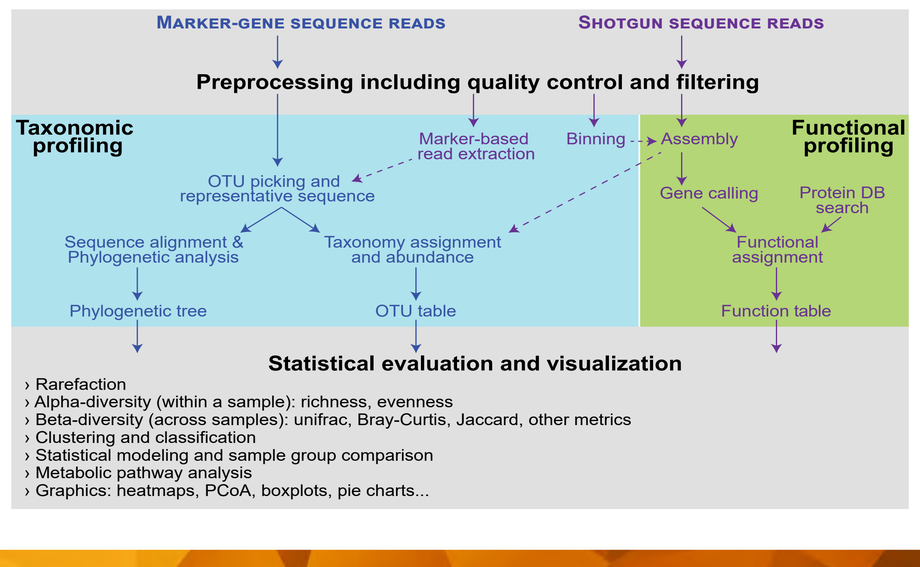
2. Data generation according to scientific question



		Sequence-based methods: Composition: Detection	Sequence-based methods: Composition: Quantification	Sequence-based methods: Function
Targeted methods	few taxa known beforehand	Fluorescence *in situ* hybridization	RT qPCR of taxon-specific genes	
	many known taxa * and unknown	Phylogenetic microarray	MARKER-GENE SEQUENCING*	
Non-targeted methods	many known and unknown taxa		SHOTGUN SEQUENCING	

3. Data analysis

Taxonomic profiling | **Functional profiling**

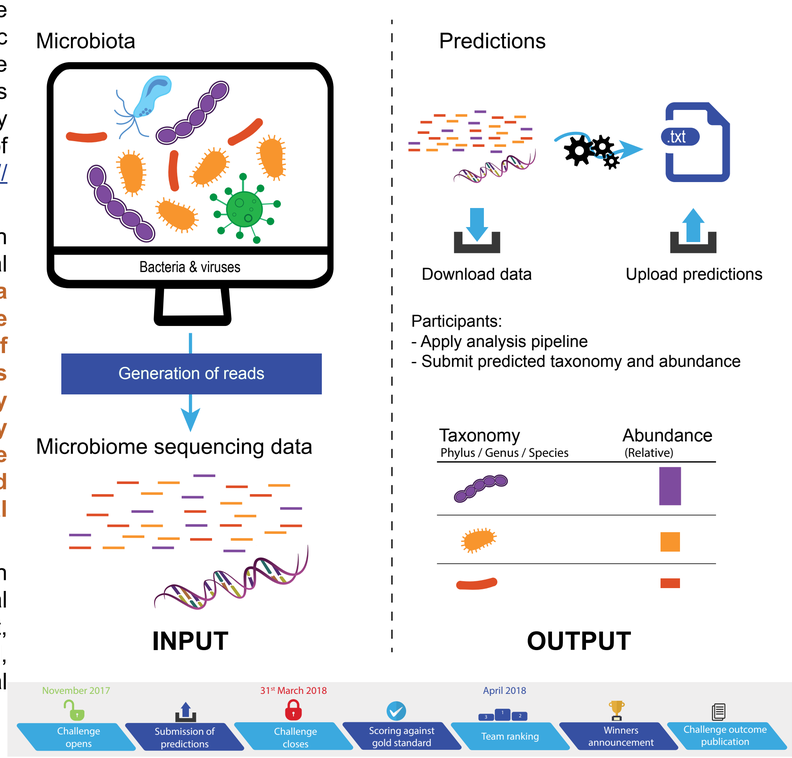


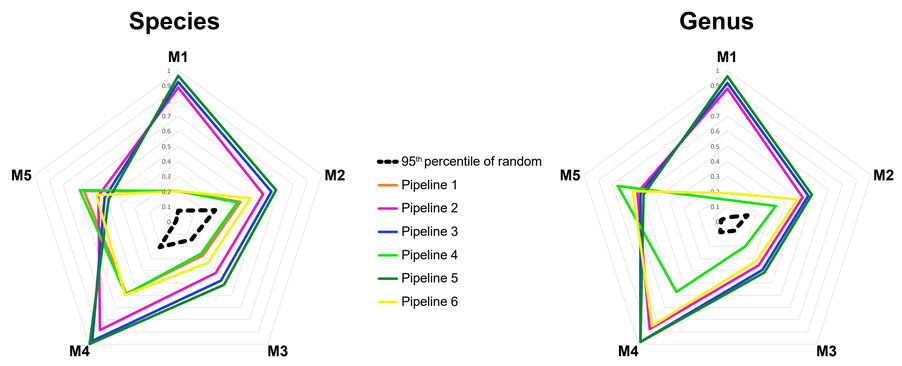
sbv IMPROVER Microbiomics Challenge

An accurate analysis of microbiome sequencing data (e.g., accurate taxonomic assignment and relative abundance estimates) relies on computational methods that have been scrutinized, partially, by initiatives such as Critical Assessment of Metagenome Interpretation (CAMI) (http://www.cami-challenge.org/).

To build and expand upon what has been done by CAMI, namely assessing individual steps of the workflow, **the "Microbiota composition prediction" challenge aims to assess the performance of metagenomics computational analysis pipeline(s) as a whole objectively (i.e., from quality control to taxonomy profiling), the end result being the recovery of relative abundance and taxonomy assignment of bacterial communities**.

Participants are provided with shotgun DNA sequencing data for several microbiome samples and asked to predict, at the phylum, genus, and species level, the relative abundance of bacterial communities present in each sample.





Example of results obtained by challenge organizers when trying to solve the challenge with different pipelines.

M1 to M5 represent different scoring metrics that are used to evaluate the performance of the methods.

It is already possible to see that different methods have different strengths. The outcome of the challenge with more diverse methods and parameters promise very interesting results.

References

Meyer, P., L. G. Alexopoulos, T. Bonk, A. Califano, C. R. Cho, A. de la Fuente, D. de Graaf, A. J. Hartemink, J. Hoeng, N. V. Ivanov, H. Koeppl, R. Linding, D. Marbach, R. Norel, M. C. Peitsch, J. J. Rice, A. Royyuru, F. Schacherer, J. Sprengel, K. Stolle, D. Vitkup and G. Stolovitzky (2011). "Verification of systems biology research in the age of collaborative competition." Nat Biotechnol 29(9): 811-815.

Poussin, C., V. Belcastro, F. Martin, S. Boue, M. C. Peitsch and J. Hoeng (2017). "Crowd-Sourced Verification of Computational Methods and Data in Systems Toxicology: A Case Study with a Heat-Not-Burn Candidate Modified Risk Tobacco Product." Chem Res Toxicol 30(4): 934-945.

Poussin, C., C. Mathis, L. G. Alexopoulos, D. E. Messinis, R. H. Dulize, V. Belcastro, I. N. Melas, T. Sakellaropoulos, K. Rhrissorrakrai, E. Bilal, P. Meyer, M. Talikka, S. Boue, R. Norel, J. J. Rice, G. Stolovitzky, N. V. Ivanov, M. C. Peitsch and J. Hoeng (2014). "The species translation challenge-a systems biology perspective on human and rat bronchial epithelial cells." Sci Data 1: 140009.

sbv IMPROVER project team, S. Boue, B. Fields, J. Hoeng, J. Park, M. C. Peitsch, W. K. Schlage, M. Talikka, P. Challenge Best, I. Binenbaum, V. Bondarenko, O. V. Bulgakov, V. Cherkasova, N. Diaz-Diaz, L. Fedorova, S. Guryanova, J. Guzova, G. Igorevna Koroleva, E. Kozhemyakina, R. Kumar, N. Lavid, Q. Lu, S. Menon, Y. Ouliel, S. C. Peterson, A. Prokhorov, E. Sanders, S. Schrier, G. Schwaitzer Neta, I. Shvydchenko, A. Tallam, G. Villa-Fombuena, J. Wu, I. Yudkevich and M. Zelikman (2015). "Enhancement of COPD biological networks using a web-based collaboration interface." F1000Res 4: 32.

E. Scotti, S. Boué, G. L. Sasso, F. Zanetti, V. Belcastro, C. Poussin, N. Sierro, J. Battey, A. Gimalac, N. V. Ivanov and J. Hoeng (2017). "Exploring the microbiome in health and disease:Implications for toxicology." Toxicology Research and Application 1: 2397847317741884.

Tarca, A. L., M. Lauria, M. Unger, E. Bilal, S. Boue, K. Kumar Dey, J. Hoeng, H. Koeppl, F. Martin, P. Meyer, P. Nandy, R. Norel, M. Peitsch, J. J. Rice, R. Romero, G. Stolovitzky, M. Talikka, Y. Xiang, C. Zechner and I. D. Collaborators (2013). "Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge." Bioinformatics 29(22): 2892-2899.

PMI SCIENCE
PHILIP MORRIS INTERNATIONAL

Competing Financial Interest

The research described in this poster was sponsored by Philip Morris Products SA