![Front Line Genomics logo — Unzipping genes for the good of humanity]

**Date:** 08 September 2016

**Headline:** Could transcriptomics predict exposure to toxicants?

**Sub-head:** An interview with Carine Poussin PhD, Philip Morris International on the sbv IMPROVER Systems Toxicology Computational Challenge

**Byline:** Liz Harley

**Article online at:** http://www.frontlinegenomics.com/press-release/7433/transcriptomics-predict-exposure-toxicants/
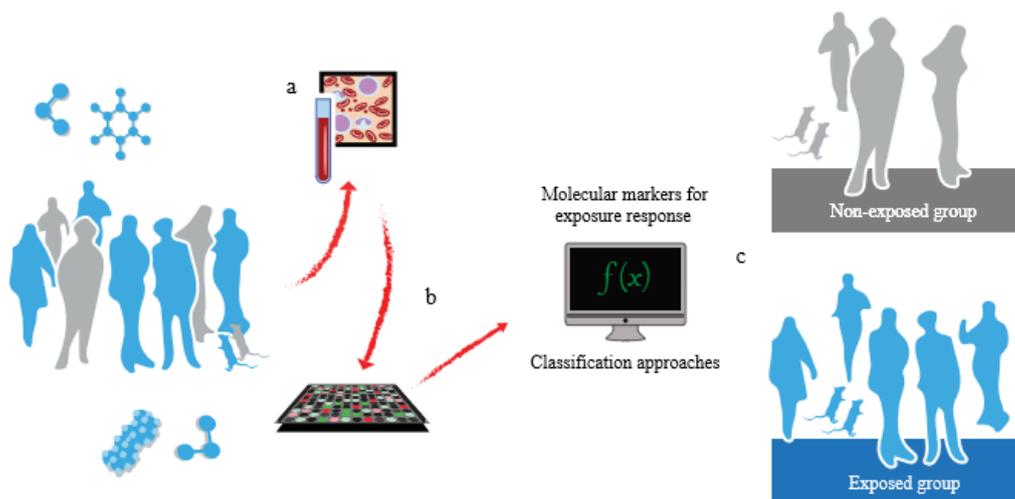


**Can you please give an overview of the recent sbv IMPROVER Systems Toxicology Computational Challenge? How many scientists registered for the challenge?**

The Systems Toxicology Computational Challenge was designed to explore the extent to which transcriptomics information present in the blood can be used to predict whether people have been exposed or not exposed to specific toxicants. Provided with blood transcriptomics datasets, participants were asked to solve two tasks. Firstly, they were asked to derive predictive classification models that would distinguish current tobacco smokers from non-current smokers (prediction of smoking exposure status). Secondly, they were asked to discriminate the non-current smokers as former smokers and never smokers (prediction of cessation status).
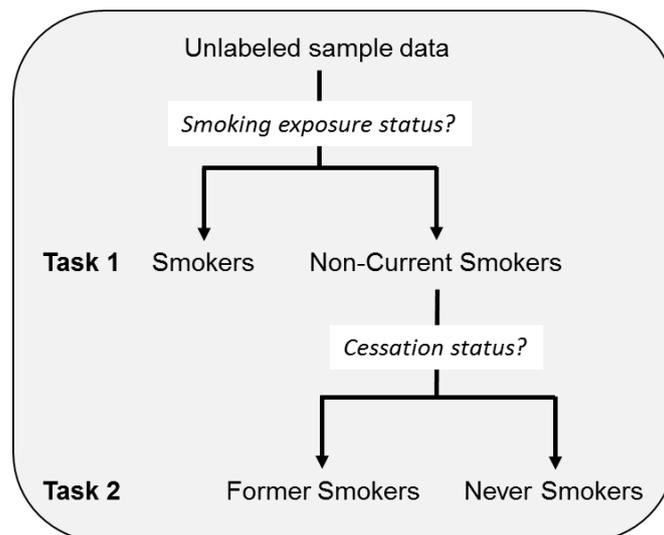
The challenge included two independent sub-challenges, each aiming to address the two tasks above within a different context. Sub-challenge one explored whether gene expression changes in human blood are sufficiently informative to predict smoking exposure or cessation status (human data only). The second investigated the issue of species translatability, asking participants to identify species-independent blood markers from in vivo rodent studies that can then be applied to clinical blood samples in order to assess exposure / cessation status in humans. Anonymized participants' submissions were scored against a gold-standard corresponding to the true class label of samples. Final results and team rankings were reviewed and approved by an independent expert scoring review panel.

The Systems Toxicology Computational Challenge was open to anyone working in computational sciences who develops predictive modelling techniques. A total of 135 scientists from around the world registered to take part and were free to use their own computational techniques to make their predictions.

**Why is it difficult to identify relevant markers in blood after chemical exposure?**

Blood is a complex tissue to analyze, primarily due to the many different cell sub-populations it contains. Molecular changes brought about by exposure to a toxicant may involve a complex interplay of a sub-set of the chemicals present in the toxicant itself, molecules produced by the exposed organ (e.g., the lungs or the gut), and chemical-derived metabolites.

Furthermore, the real-world application of models based on blood markers for predictive classification of individuals is also uniquely challenging. As well as the problem of identifying relevant markers in blood after chemical exposure, the difficulty also resides in the low success of correct classification when predictive models are applied on new individual blood samples, and the translation of these techniques into practical ready-to-use tools.

**What different computational techniques were used by challenge participants and what level of accuracy was achieved?**

Many different computational techniques were proposed, with the best-performing teams using random forest, linear discriminant analysis, partial least square discriminant analysis and logistic regression machine learning methods. Best-performers achieved accuracy of up to 95% in distinguishing current tobacco smokers from non-smokers. Predicting whether non-smokers were former smokers or never smokers was more challenging, suggesting that these two groups are likely to have similar gene expression profiles. Exposure response markers identified included a core gene subset that was highly consistent across teams.

**In what way have the results of the Systems Toxicology Computational Challenge improved our understanding of what is necessary to reach higher levels of predictability?**

The Systems Toxicology Computational Challenge has demonstrated what is necessary to reach high levels of predictability using techniques that are potentially suitable for ready-to-use diagnostic tools. The challenge rules stipulated that the models proposed must be capable of being applied to new, individual blood samples without the need for adjustments (i.e., they had to be inductive). This decision was taken following insights gleaned from a previous sbv IMPROVER challenge – the Diagnostic Signature Challenge – which assessed the extent to which markers for four distinct diseases could be extracted from data held in public repositories (training datasets), and then used to make diagnostic predictions on unrelated datasets (test datasets). An interesting aspect of this earlier challenge was that most of the models developed were transductive, i.e., they relied on processing both training and test datasets within the same model, with class predictions then being made on subjects from the test datasets.

When it comes to real-world application, transductive approaches to prediction are limited since they cannot reliably be generalized and independently applied to new datasets. All participants in the Systems Toxicology Computational Challenge succeeded in the development of inductive classification models.

In more general terms, the Systems Toxicology Challenge provided the opportunity for the global scientific community to vigorously and objectively test their independent predictive methodologies on high-quality, large scale datasets. This crowdsourcing approach has led to transparency and confirmation of the relative success of different predictive techniques, in both humans and across species.

**Were you surprised by participants' success in developing predictive models?**

As a computational biologist I was confident in the ability of modern techniques, if properly applied, to accurately predict exposure status from blood transcriptomics data. However, it was fantastic to see the variety of approaches used, the high level of predictability achieved and the identification of a core gene subset that was consistent across teams.

Is it likely that the techniques the participants put forward could be applied to make predictions on other toxicants or external stimuli?

Yes. While the Systems Toxicology Computational Challenge asked participants to make predictions on smoking status, the techniques that participants put forward could in theory be applied to make predictions on exposure to any toxicant or external stimuli. Exposure to all external toxicants can induce molecular changes in human blood, and being able to identify exposure status from blood samples (which are easily accessible) has valuable implications for the toxicological risk-assessment of chemicals, drugs and consumer products, as well as for diagnostics.

**Could any of the techniques be combined to improve predictability even further?**

Aggregation of the best-performing methodologies has the potential to improve predictability further, and add to the confidence we can have in using them. However, for the second task – determining whether non-smokers are former smokers or never smokers – we would not expect to see any significant improvement in predictability since former smokers were frequently misclassified across all methodologies, suggesting that non-smokers have similar gene expression profiles regardless of whether they previously smoked and then quit for a certain period of time, or never smoked at all.

**What do you think the future holds for using transcriptomics information to predict exposure to toxicants?**

The Systems Toxicology Computational Challenge has demonstrated how 'omics data collected from blood samples can be used to predict toxicant exposure status. Best-performers achieved near perfect prediction in discriminating smokers from non-smokers. The computational techniques explored in the challenge, as well as key learnings, should be of great interest for industries such as pharmaceuticals and biotechnology. We would like to see these techniques investigated as proof of concept tools for product development in diagnostics, toxicological risk-assessment and personalized medicine.

**Do sbv IMPROVER plan to run any more challenges in the near future?**

Yes. The Systems Toxicology Computational Challenge is the fourth challenge to run under the sbv IMPROVER umbrella and the project continues. On 23-24 September 2016, an sbv IMPROVER Datathon event will be held in Singapore, bringing together data scientists and bioinformaticians to interpret data in an open, collaborative forum. Registration for the Datathon is now open. A further sbv IMPROVER challenge, which we hope to launch next year, is currently under development.

**Where can readers find more information?**

Full details of the sbv IMPROVER project are available at [www.sbvimprover.com](http://www.sbvimprover.com). Specific details on the upcoming Datathon, including the registration portal, can be accessed via the [dedicated event website](http://).

**About Carine Poussin PhD**

Trained as computational biologist, Dr Carine Poussin is part of the scientific team responsible for the development of sbv IMPROVER computational challenges at Philip Morris International. Dr Carine Poussin is also leading various innovative computational and experimental research and systems toxicology-based product assessment projects. Working for many years in academia and industry, Dr Carine Poussin has acquired broad experience in computational developments, data analysis and interpretation in diverse research areas ranging from cardiovascular, lung, immune and metabolic diseases to nutrition and toxicology.

### ###

### sbv IMPROVER Systems Toxicology Computational Challenge
### – Best-Performing Teams –

| Sub-Challenge 1 | Team Members |
|---|---|
| 1st | Adi L. Tarca, Associate Professor, Wayne State University, School of Medicine, USA <br><br> Prof Roberto Romero, Wayne State University, School of Medicine, USA |
| 2nd | Anonymous Entry |
| 3rd | Xiaofeng Gong, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China <br><br> Wenxin Yang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China <br><br> Zhongqu Duan, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China <br><br> Peixuan Wang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China |

| | Hao Yang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China |
| | Chenfang Zhang, Shanghai Jiao Tong University-Yale Joint Center for Biostatistics, China |

| Sub-Challenge 2 | Team Members |
|---|---|
| 1st | Ömer Sinan Saraç Associate Professor, Istanbul Technical University, Turkey |
| | İsmail Bilgen, Istanbul Technical University, Turkey |
| | Ali Tuğrul Balcı, Istanbul Technical University, Turkey |
| 2nd (joint) | Adi L. Tarca, Associate Professor, Wayne State University, School of Medicine, USA |
| | Prof Roberto Romero, Wayne State University, School of Medicine, USA |
| 2nd (joint) | Dr Sandeep Kumar Dhanda, La Jolla Institute for Allergy & Immunology, USA |
| | Dr Rahul Kumar, London, UK |

- - - - - - - - -

*Homepage Screengrab ([www.frontlinegenomics.com](www.frontlinegenomics.com), as at 08 September 2016):*



-- ENDS --