Computational Toxicology xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status

Vincenzo Belcastro [a,*,1], Carine Poussin [a,*,1], Yang Xiang [a], Maurizio Giordano [o], Kumar Parijat Tripathi [o], Akash Boda [a], Ali Tugrul Balci [l,3], Ismail Bilgen [l,3], Sandeep Kumar Dhanda [n,3], Zhongqu Duan [i,k,3], Xiaofeng Gong [i,k,3], Rahul Kumar [m,3], Roberto Romero [d,e,f,g,h,3], Omer Sinan Sarac [l,3], Adi L. Tarca [b,c,3], Peixuan Wang [i,k,3], Hao Yang [i,j,3], Wenxin Yang [i,j,3], Chenfang Zhang [i,k,3], Stéphanie Boué [a], Mario Rosario Guarracino [o], Florian Martin [a], Manuel C. Peitsch [a], Julia Hoeng [a]

[a] PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchatel, Switzerland[2]
[b] Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI, USA
[c] Department of Computer Science, Wayne State University College of Engineering, Detroit, MI, USA
[d] Perinatology Research Branch, NICHD/NIH/DHHS, Bethesda, MD, USA
[e] Perinatology Research Branch, NICHD/NIH/DHHS, Detroit, MI 48201, USA
[f] Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, MI 48109, USA
[g] Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI 48825, USA
[h] Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA
[i] SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China
[j] School of Mathematics Sciences, Shanghai Jiao Tong University, Shanghai, China
[k] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China
[l] Istanbul Technical University, Istanbul, Turkey
[m] Institute of Microbial Technology, Sector 39A, Chandigarh 160036, India
[n] La Jolla Institute for Allergy and Immunology, 9420, Athena Circle, La Jolla, CA 92037, USA
[o] Istituto di Calcolo e Reti ad Alte Prestazioni CNR, Via P. Castellino, 111 80131 Napoli, Italy

ARTICLE INFO

Article history:
Received 20 April 2017
Received in revised form 23 June 2017
Accepted 12 July 2017
Available online xxxx

Keywords:
Systems toxicology
Computational challenge
Gene signature
Smoking biomarker
Blood biomarkers

ABSTRACT

Cigarette smoking entails chronic exposure to a mixture of harmful chemicals that trigger molecular changes over time, and is known to increase the risk of developing diseases. Risk assessment in the context of 21st century toxicology relies on the elucidation of mechanisms of toxicity and the identification of exposure response markers, usually from high-throughput data, using advanced computational methodologies.

The sbv IMPROVER Systems Toxicology computational challenge (Fall 2015-Spring 2016) aimed to evaluate whether robust and sparse (≤40 genes) human (sub-challenge 1, SC1) and species-independent (sub-challenge 2, SC2) exposure response markers (so called gene signatures) could be extracted from human and mouse blood transcriptomics data of current (S), former (FS) and never (NS) smoke-exposed subjects as predictors of smoking and cessation status. Best-performing computational methods were identified by scoring anonymized participants' predictions.

Worldwide participation resulted in 12 (SC1) and six (SC2) final submissions qualified for scoring. The results showed that blood gene expression data were informative to predict smoking exposure (i.e. discriminating smoker versus never or former smokers) status in human and across species with a high

* Corresponding authors.
E-mail addresses: vincenzo.belcastro@contracted.pmi.com (V. Belcastro), carine.poussin@pmi.com (C. Poussin), yang.xiang@pmi.com (Y. Xiang), maurizio.giordano@cnr.it (M. Giordano), parijat24@gmail.com (K.P. Tripathi), akash.boda@uqconnect.edu.au (A. Boda), alitugrulbalci@gmail.com (A.T. Balci), ibilgen@itu.edu.tr (I. Bilgen), mdusdhanda@gmail.com (S.K. Dhanda), zhqduan@sjtu.edu.cn (Z. Duan), ikergong@gmail.com (X. Gong), rahulk.aiims@gmail.com (R. Kumar), roberto.romero@wayne.edu (R. Romero), ssarac@itu.edu.tr (O.S. Sarac), adi@wayne.edu (A.L. Tarca), peixuan_wang@qq.com (P. Wang), eric_yang09@126.com (H. Yang), ywxsjtu@sjtu.edu.cn (W. Yang), cfzhang2015@sjtu.edu.cn (C. Zhang), stephanie.boue@pmi.com (S. Boué), mario.guarracino@na.icar.cnr.it (M.R. Guarracino), florian.martin@pmi.com (F. Martin), manuel.peitsch@pmi.com (M.C. Peitsch), julia.hoeng@pmi.com (J. Hoeng).

[1] Equal contribution.
[2] Part of Philip Morris International group of companies.
[3] Challenge best performers.

http://dx.doi.org/10.1016/j.comtox.2017.07.004

level of accuracy. By contrast, the prediction of cessation status (i.e. distinguishing FS from NS) remained challenging, as reflected by lower classification performances. Participants successfully developed inductive predictive models and extracted human and species-independent gene signatures, including genes with high consensus across teams. Post-challenge analyses highlighted "feature selection" as a key step in the process of building a classifier and confirmed the importance of testing a gene signature in independent cohorts to ensure the generalized applicability of a predictive model at a population-based level.

In conclusion, the Systems Toxicology challenge demonstrated the feasibility of extracting a consistent blood-based smoke exposure response gene signature and further stressed the importance of independent and unbiased data and method evaluations to provide confidence in systems toxicology-based scientific conclusions.

Introduction

Holy grail in systems toxicology: can specific markers of exposure response to chemical(s) be identified in blood?

Humans are constantly exposed to individual or mixtures of chemicals (e.g., cigarette smoke, pollutants, pesticides, and other chemicals) that may have effects on their cells. If chemicals are harmful and/or cumulative doses of chemicals exceed a threshold limit value, exposure can lead to cellular/tissue damage and dysfunction, which in turn can increase risk of disease development. Hence, identification of specific markers elicited in response to (specific) chemicals is important to assess the exposure status of subjects and to associate exposure with toxicity outcomes. The appropriate combination of identified markers constitutes a specific exposure response fingerprint or signature discriminating exposed and non-exposed subjects. Such a signature may also distinguish formerly-exposed and never-exposed subjects. Exogenous chemicals (e.g., lead), chemical-derived metabolites, and endogenous molecules produced by primarily exposed organs (e.g., lung, gut) can pass into the blood stream and may induce molecular changes in blood cells [1]. Therefore, investigating whether specific markers in response to chemical exposure can be identified in blood cells may be highly valuable for monitoring chemical exposure [2,3]. Interestingly, new 'omics' technologies (e.g., genomics, transcriptomics, proteomics, metabolomics, lipidomics) can be applied to toxicity testing in order to increase efficiency and provide a more data- and system-driven approach to exposure response assessment [3,4].

Gene signature-based classification models for biological/clinical status prediction

Transcriptomics-based technologies enable biological mechanistic insights to be gained by measuring whole-genome gene expression levels. Transcriptomics data have also been used extensively to develop classification models predictive of disease diagnosis or prognosis, tumor subtyping, adverse drug response, and therapeutic outcome [5–8]. Gene signatures are generally derived from disease-relevant tissues such as liver, lung, and tumors. However, blood can be collected easily for diagnostics (minimally invasive) and only small quantities are necessary for transcriptomics profiling. Therefore, more and more investigations have used blood samples to identify gene signatures that may be leveraged for the development of tests such as In Vitro Diagnostic Multivariate Index Assays [9,10]. The real-world application of gene signature-based classification models as reliable tools for predictive medicine is still limited [11]. This is mainly because (i) it is difficult to identify robust and sufficiently sparse signatures for the development of ready-to-use diagnostic tools and (ii) the way models are built often leads to poor predictive performance when applied to new individual samples (e.g., lack of validation in independent cohorts to test robustness and generalized applicability in populations, or the use of transductive method-based models [12]).

The systems biology verification Industrial Methodology for PROcess VErification in Research (sbv IMPROVER [13]; https://sbvimprover.com) project aims to verify methods and data in systems biology/toxicology using double-blind performance assessment. Over the past six years, sbv IMPROVER organized crowd-sourced challenges covering a broad range of scientific questions [14–18]. The first one titled Diagnostic Signature Challenge in 2012 was designed to assess to what extent models trained on transcriptomics data available in public repositories could predict the diagnosis of individual subjects in unrelated datasets for four disease types [18]. Many of the classification models proposed were transductive (i.e., training and test sets are processed together and prediction model solely applies to this specific test set) rather than inductive (i.e., the signature model is applied to a single new sample without retraining), which may lead to poor classification on a new single patient sample and may be impracticable for real-world application. These limitations were considered in the design and constraint of new classification problems in our latest computational challenge open to the scientific community and described below.

Application of omics-based classification to toxicogenomics using blood as surrogate tissue: prediction of tobacco smoke exposure and cessation

In liver and pulmonary toxicity studies, gene signatures have been identified successfully in blood showing (i) capability to predict exposure and toxicity to chemicals such as acetaminophen in liver (drug-induced liver injury) or crystalline silica in lung; (ii) superior sensitivity as predictors of toxicity compared with the classical toxicity markers in rats; and (iii) to some extent, similarities in pathways and functions that are perturbed in primary tissue and blood [3]. These findings, in addition to its easy access, make blood highly relevant as a surrogate to identify gene expression-based signatures as specific markers for toxicological evaluation and risk assessment. Smoking is a major risk factor for the development of various diseases (e.g., cardiovascular and lung diseases) [19]. Smokers are exposed to a mixture of thousands of chemical constituents when cigarette smoke is inhaled. Among them, some constituents or their metabolites that pass into the blood circulation elicit systemic effects distal from the lungs, the primary site of exposure. For example, changes in gene expression in circulating peripheral blood cells are associated with several systemic immune and inflammatory-related disorders [20,21]. Smoking cessation has been shown to revert some cigarette smoke-induced functional and molecular changes back to non-smoker levels or intermediate levels depending on the subject's smoking history (e.g., smoking duration, consumption) and

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

cessation period [22–24]. Therefore, the identification of specific markers of response to smoking or cessation in whole blood cells may be an important way to monitor the exposure status of an individual subject. In general, blood-based signatures for smoking exposure that have been reported in the literature share very few genes [25–27]. Most pre-clinical toxicological *in vivo* studies are conducted in rodents, adding a degree of complexity when applying the results to humans. This raises the question of the translatability of blood-based exposure response gene signatures between human and rodents that was also addressed in the challenge.

The Systems Toxicology computational challenge

The latest sbv IMPROVER computational challenge in the scope of this manuscript and titled "Systems Toxicology" (from November 2015 to May 2016) had the objective to assess whether gene expression data from blood cells were sufficiently informative to extract specific blood response markers (i.e. signature) to predict smoking exposure and cessation status in human (sub-challenge 1, SC1) and across species for (sub-challenge 2, SC2) translational toxicology (https://sbvimprover.com/challenge-4/the-computational-challenge; Fig.1a). Participants were asked to develop blood-based gene signature classification models that could (i) discriminate smoking exposure status, i.e. separate smoke-exposed and non-current smoke-exposed subjects, and (ii) among non-current smoke-exposed subjects, distinguish formerly and never smoke-exposed subjects (cessation status) (Fig.1b).

The challenge rules constrained participants to develop inductive rather than transductive predictive classification models that could identify gene signatures that did not exceed 40 genes (sparsity). The Affymetrix gene expression datasets provided for the challenge included blood samples from smoker/smoke-exposed, former smoker/formerly smoke-exposed, and never smoker/never smoke-exposed groups of subjects from clinical (human) and *in vivo* mouse studies. The participants received class labels and data for the training dataset and also had the freedom to use additional public and/or private gene expression data to train their models. Then, the participants were asked to apply their trained predictive classification models directly on new data from independent studies (test dataset) and to provide confidence levels that a blood sample belonged to one class or the other. After closure of the challenge, anonymized participants' predictions were scored using pre-defined metrics for performance assessment and best-performing teams were identified. The present manuscript summarizes the results and learnings from the Systems Toxicology computational challenge.

Materials and methods

Study population and designs

Whole blood samples were acquired in the context of clinical and *in vivo* mouse studies conducted by Philip Morris International, or from a Biobank repository. Populations and designs for the different studies are described below. Whole blood was collected in PAXgene™ tubes and frozen at −80 °C until it was further processed to generate transcriptomics data. Details of the different studies described hereafter are available in Supplementary table 1.

QASMC study (dataset H1)

The Queen Ann Street Medical Center (QASMC) study is a clinical case–control study [28] that was conducted between July 2011 and December 2012 at The Heart and Lung Centre (London, UK), after approval from the National Health Service (NHS) Black Country Ethics Committee and in strict compliance with the International Conference on Harmonization-Good Clinical Practice (ICH-GCP) guidelines. The study was registered at ClinicalTrials.gov with the identifier NCT01780298. The QASMC study aimed to identify biomarker(s) that would enable differentiation between smokers with chronic obstructive pulmonary disease (COPD) (i.e., cigarette smoke with a ≥10 pack/year smoking history and COPD disease classified as GOLD Stage 1 or 2) and three comparative

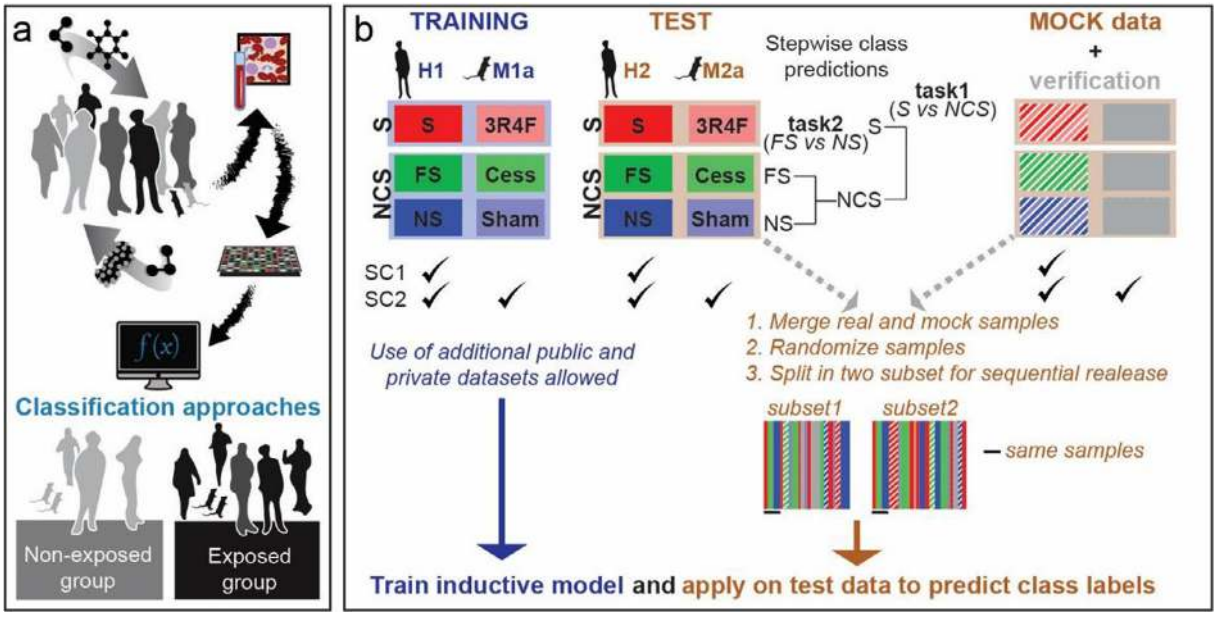

Fig. 1. Overview of the Systems Toxicology computational challenge. (a) Human and mouse blood samples were collected from smokers/3R4F-exposed (S/3R4F) and non-current smokers/not-exposed-to-3R4F (NCS) (mouse: exposed and non-exposed) and gene expression was measured. Classification approaches were developed by the participants to identify exposed and non-exposed subjects. (b) Human and mouse samples were divided into training (H1 and M1a) and test (H2 and M2a) datasets. Training datasets and class labels were released to allow participant to train their models. Test datasets (including mock samples) were released in two subsets. Participants were asked to provide their predictions on the first subset before the second subset was released. Participants had to apply their models to assess the class labels for the samples in the test set.

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

groups of matched subjects (matched by ethnicity, sex, and age (within 5 years) with the recruited COPD subjects): smokers (S), former smokers (FS), and never smokers (NS). All smoking subjects (S and FS) had a smoking history of at least 10 pack-years. FS quit smoking at least 1 year prior to sampling (~78% of FS have stopped for more than 5 years). Sixty subjects in each group were enrolled (~240 subjects in total:~60 S with COPD,~60 S, 60 FS, and 60 NS). The 240 patients included males (58%) and females (42%) aged between 40 and 70 years. The microarray data are available in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-5278.

BLD-SMK-01 (dataset H2)

The transcriptomics dataset BLD-SMK-01 (H2) was produced from PAXgene™ blood samples obtained from a banked repository (BioServe Biotechnologies Ltd., Beltsville, MD, USA) based on defined inclusion/exclusion criteria as follows. At the time of sampling, the subjects were between 23 and 65 years of age. Subjects with a disease history and those taking prescription medications were excluded. Smokers had smoked at least 10 cigarettes daily for at least three years. Former smokers had ceased smoking at least two years prior to sampling and before quitting had smoked at least 10 cigarettes daily for at least three years. Smokers and never smokers were matched by age and sex, while former smokers could not be properly matched due to the lower number of samples available for this group. The microarray data are available in the ArrayExpress database under accession number E-MTAB-5279.

ZRHR-reduced exposure (REX)C-03-EU and -04-JP studies (datasets H3 and H4)

The REXC-03-EU and C-04-JP studies were randomized, controlled, open-label, 3-arm parallel group, and single-center studies (Supplementary Fig. 1a). These studies were performed to demonstrate reductions in exposure to selected smoke constituents for healthy subjects switching to the tobacco heating system THS2.2 (Switch) corresponding to a candidate modified-risk tobacco product (MRTP), or smoking abstinence (cessation or Cess), compared with continuing to use conventional cigarettes (considered as smokers), for 5 days in confinement [29,30]. The studies were conducted in Europe (*C-03-EU*) and Japan (*C-04-JP*) and were registered at ClinicalTrials.gov with the identifiers NCT01959932 and NCT01970982, respectively. The microarray data are available in the ArrayExpress database under accession numbers E-MTAB-5332 (C-03-EU) and E-MTAB-5333 (C-04-JP).

Mouse C57Bl6-pMRTP-SW inhalation study (dataset M1a/b)

A 7-month inhalation study was conducted with female C57BL/6 mice [24]. The study design (Supplementary Fig.1b) included five groups of mice that were exposed to: (1) cigarette smoke from a reference cigarette (3R4F), (2) air after 2-month exposure to 3R4F (Cess), (3) air only (Sham), (4) mainstream aerosol from a potential MRTP (pMRTP), or (5) switched to a pMRTP after 2-month exposure to 3R4F (Switch) [24]. Data from the latter two groups were provided to the challenge participants for verification purposes and results associated with these data are discussed in a separate manuscript [31]. For each group, blood samples were collected from 7 to 10 animals at 2, 3, 5 and 7 months. Samples from the 3R4F, pMRTP, and Sham groups were also collected at 4 months. The microarray data are available in the ArrayExpress database under accession number E-MTAB-5281.

Mouse Apoe$^{-/-}$ -THS2.2-SW inhalation study (dataset M2a/b)

An 8-month inhalation study was conducted with female Apoe$^{-/-}$ mice randomized into different groups [32]. The study design (Supplementary Fig. 1c) included five groups of mice that have been exposed to (1) cigarette smoke from 3R4F, (2) air for 6 months after 2-month exposure to 3R4F (Cess), (3) air (Sham) for up to 8 months, (4) mainstream aerosol from THS2.2, at nicotine levels matched to those of 3R4F for up to 8 months, or (5) THS2.2 for 6 months after 2-month exposure to 3R4F (switch). Data from these two latest groups were provided to the challenge participants for verification purposes and results associated with these data are discussed in a separate manuscript [31]. For each group, blood was sampled at 1, 2, 3, 6, and 8 months. The microarray data are available in the ArrayExpress database under accession number E-MTAB-5280.

Both potential MRTP and THS2.2 (the candidate MRTP) are heat-not-burn tobacco-based technologies. The tobacco heating system (THS)2.2, uses an electrically heated system [33], and pMRTP uses a fast-lighting carbon tip as heat source [24].

Transcriptomics data generation and processing

RNA isolation from human blood samples

For each clinical and *in vivo* study, the samples were randomized prior to RNA extraction. For the clinical studies, total RNA was isolated using a PAXgene™ Blood miRNA Kit (catalog number 763134; Qiagen, Venlo, The Netherlands) according to the manufacturer's instructions. For the *in vivo* mouse studies, total RNA was isolated using a RNeasy Protect Animal Blood Kit (catalog number 73224; Qiagen, Venlo, The Netherlands) according to the manufacturer's instructions. The concentration and purity of the RNA samples were determined using a UV spectrophotometer (NanoDrop® 1000 or Nanodrop 8000; Thermo Fisher Scientific, Waltham, MA, USA) by measuring the absorbance at 230, 260, and 280 nm. RNA integrity was further checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Only RNA samples with a RNA integrity number (RIN) >6 were processed for further analysis.

RNA preparation and hybridization on Affymetrix chip

Targets were prepared from 80 ng of RNA using the Ovation® Whole Blood Reagent and Ovation RNA Amplification System V2 (NuGEN, AC Leek, The Netherlands). The quantity of cDNA was measured with a SpectraMax® 384Plus microplate reader (Molecular Devices, Sunnyvale, CA, USA). The cDNA quality was determined by assessing the size of the unfragmented cDNA using a Fragment Analyzer (Advanced Analytical, Ankeny, IA, USA). The size distribution of the final fragmented and biotinylated product was also monitored using electropherograms on an Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). After fragmentation and labeling the cDNA fragments for human and mouse were hybridized on a GeneChip® Human Genome U133 Plus 2.0 Array or GeneChip® Mouse Genome 430 2.0 Array (Affymetrix), respectively, according to the manufacturer's guidelines.

For the QASMC study (H1 dataset), target preparation from blood samples and hybridization on a GeneChip® Human Genome U133 Plus 2.0 Array (Affymetrix) were performed by AROS Applied Biotechnology AS (Aarhus, Denmark).

Raw data processing and QC

Raw data (CEL files) from each dataset were processed and normalized in the R environment (v3.1.2, [34]) using frozen Robust Microarray Analysis, fRMA v1.18 [35]. Frozen parameter vectors for mouse (mouse4302frmavecs v1.3.0, [36]) and human (hgu133plus2frmavecs v1.3.0, [37]) were used by the frma functions. The custom brainarray cdf files for mouse (mouse4302mmentrezgcdf v16.0.0) and human (hgu133plus2hsentrezgcdf v16.0.0 [38]) were used for Affymetrix probe-to-Entrez Gene ID mapping, resulting in one probe set for one gene (for details, see Supplementary information). Normalized-unscaled standard error (NUSE), relative log

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

expression (RLE), median absolute value RLE (MARLE) and pseudo-images as well as raw images plot were generated for quality check of the data.

Differential gene expression analysis

Differential expression for contrasts of interest (smoker vs non-current smoker) was determined by linear modeling using the limma R package v3.22.1 [39].

The Systems Toxicology challenge

Goals and rules

Participants were asked to develop robust and sparse human (sub-challenge 1, SC1) and species-independent (sub-challenge 2, SC2) blood-based gene signature classification models to discriminate between smoke-exposed and non-currently smoke-exposed subjects (task 1, Fig.1b), and subsequently (ii) to classify non-currently smoke-exposed subjects, as former and never smoke-exposed subjects (task 2, Fig.1b). As a first constraint, predictive models were requested to be inductive, as opposed to transductive [12], with the ability to predict the class of a single new blood sample without the need to retrain/refine the model or use semi-supervised approach(es) combining training and test datasets to predict sample class. Participants had the opportunity to develop 3-class prediction model. As a second constraint, the length of the signatures could not exceed 40 genes.

Data preparation and release

After data processing and normalization, datasets were prepared for release to the participants. Sample names were anonymized using random generated unique names in the format "S_+10 alphanumeric characters". The metadata included the sex for each sample.

Training datasets. Participants were provided with data and class/group labels from the H1 (Clinical QASMC study) and M1a (Mouse C57Bl6-pMRTP-SW inhalation study) datasets to train their blood-based gene signature classification models (Fig.1b and Supplementary table 1). H1 was released as the training set for SC1 (224 samples, and 18,604 genes). H1 (human) and M1a (mouse) were released as the training sets for SC2 (224 and 112 samples, respectively). Only human genes with a mouse homolog (14,311 genes) were retained in the two sets, as described below. The participants had the freedom to use additional private and/or public blood-related datasets for training.

Test and verification datasets. Participants applied their trained blood-based gene signature models (see Supplementary table 2 for the full list of methods applied) on test sets of gene expression data: H2 for SC1 and H2 plus M2a (Mouse Apoe-THS2.2-SW inhalation study) for SC2. Additional datasets provided for verification purposes [31] and not used for scoring were the H3/H4 (REXC-03-EU and C-04-JP) datasets for SC1, and the H3/H4 and M1 b/M2 b datasets for SC2. For each sub-challenge, test and verification sample data were randomized and then split into two subsets released sequentially at different dates during the challenge (Fig. 1b). Samples from the different exposure groups were similarly distributed between the two subsets. To check whether participants did use inductive rather than transductive methods for class prediction, 10% of samples of subset 1 were included in subset 2. The class predictions and confidence values for those samples had to match between the two subsets. Mock sample data were added and randomized with the original sample datasets to avoid group identification using unsupervised analysis of the test dataset (e.g., clustering). Details of the preparation and release of the test and verification datasets for the challenge are available in the Supplementary information.

Human-mouse homology mapping procedure

Mouse genes were orthologized to human genes using the NCBI/HCOP mapping database [40] (download 11 Dec. 2014). HCOP aggregates orthology predictions from multiple sources. Ortholog candidates were selected on the basis of supporting evidence from these sources. HomoloGene database was selected as preferred source. For mouse genes mapping to multiple human genes with equal support, the ortholog candidate with matching gene symbol was preferred. Alternatively, the first candidate was selected. To facilitate the dataset handling, human and mouse gene expression datasets were provided with mouse gene symbols for both. A list of the orthologous genes between human and mouse that were retained for the datasets is given in Supplementary table 3.

Predictive blood-based gene signature classification models

Participants applied diverse selection methods including filter, wrapper, and embedded methods [41] to identify genes discriminative of exposure groups. Various machine learning methods were used to train blood-based gene signature classification models and estimate performances using cross-validation (Supplementary table 2).

Scoring participants' class predictions

For each sample from the test and verification datasets, participants were requested to provide a confidence value P (between 0 and 1) that the sample belonged to class 1 (e.g. smokers), and a confidence value 1-P that the sample belonged to class 2 (e.g. non-current smokers). P and 1-P were requested to be unequal. Prior to the challenge opening, the scoring strategy was defined and approved by an external and independent Scoring Review panel of experts in the field. Samples present in the test dataset, and not in the verification dataset, were used to assess team performances in each sub-challenge. Anonymized participants' class predictions were scored using the area under the precision recall (AUPR) curve and Mathews correlation coefficient (MCC) metrics [42]. The AUPR was computed as the integral between 0 and 1 of the area drawn by the precision and recall values. For MCC, a threshold of 0.5 was defined to binarize confidence values and build confusion matrices to compute MCC values that range between −1 and +1. The formulae used are reported as Supplementary information. Team performance was based on the average rank computed across metrics and tasks (task 1: smokers (S) vs non-current smokers (NCS), task 2: former smokers (FS) vs never smokers (NS). Participants incorrectly classifying smokers in task 1 were penalized in a way that these misclassified subjects were automatically placed in the wrong class for task 2 before participants were scored.

The statistical significance of the participants' predictions was verified against 10,000 randomly generated predictions. Random predictions were generated by assigning all samples from a test dataset a confidence value P ranging between 0 and 1 to be a smoker (1-P to be a non-current smoker) extracted from a uniform random distribution. Subjects with a random confidence value below 0.5 were subsequently randomly classified as former or never smoker following the same procedure. Random predictions were then scored as described above. The P-values associated with the participants' AUPR or MCC scores corresponded to the proportion of trials when the actual score was greater than a random score corresponding to the 95th percentile of scores obtained by random simulation. Scoring results and final ranking were reviewed and approved by the Scoring Review panel.

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

Post-challenge analyses

Predictions across teams and statistical distribution comparison

The "All teams" confidence values were obtained by taking the median value per subject/sample across all teams' confidence values. Confidence values were transformed to log odds, and the Wilcoxon-Mann-Whitney test was computed to compare log odds distributions between classes.

Misclassification significance

The log odds of the '1-0' outcome given by each team on all individual samples were modeled using a linear mixed model. The smoking status was included as a fixed predictor in the model while the teams were included as a random factor. The impact of smoking status on the log odds were estimated and tested for significance, and consequently on the classification error.

Crowd consensus gene signatures

Consensus human and species-independent smoking exposure and cessation gene signatures were extracted by selecting genes with at least two occurrences across the 12 and six qualified teams for SC1 and SC2, respectively.

Heatmaps of gene expression fold changes for smoking exposure and cessation consensus signatures

Gene expression fold changes were computed by subtracting the mean of normalized $\log_2$-based expression levels of the respective control groups (Fig. 5 and Supplementary Fig. 5) from each subject's normalized $\log_2$-based expression levels. Expression fold-changes of genes from the consensus signatures were visualized on heatmaps with hierarchical clustering (Euclidean as distance metric and complete method as agglomerative algorithm). Differentially expressed genes (DEGs; FDR ≤ 0.05) were determined using the limma R package version 3.22.7 [43].

Pathway/process over-representation analysis of consensus smoking exposure and cessation gene signatures

Over-representation analysis of the blood-based smoking exposure and cessation consensus gene signatures were conducted using gene set sub-collections from Broad-MSigDBv 5.1 [44] and DAVID v6.7 [45].

Performance analysis of all gene combinations from the top six teams' human-based smoking exposure consensus signature: impact of gene signature length, gene expression co-linearity level, and classification methods

The analysis included all possible combinations of genes from the consensus signature. The extraction of an 18 gene-based human smoking exposure consensus signature was limited to the top six teams (instead of the 12 qualified teams) because of limitations imposed by the computer intensive calculation required for this analysis. The 18 gene-based consensus signature in blood, which included *DSC2*, *FSTL1*, *GPR63*, *GSE1*, *GUCY1A3*, *RGL1*, *CTTNBP2*, *F2R*, *SEMA6B*, *CDKN1C*, *CLEC10A*, *GPR15*, *LINC00599*, *P2RY6*, *PID1*,

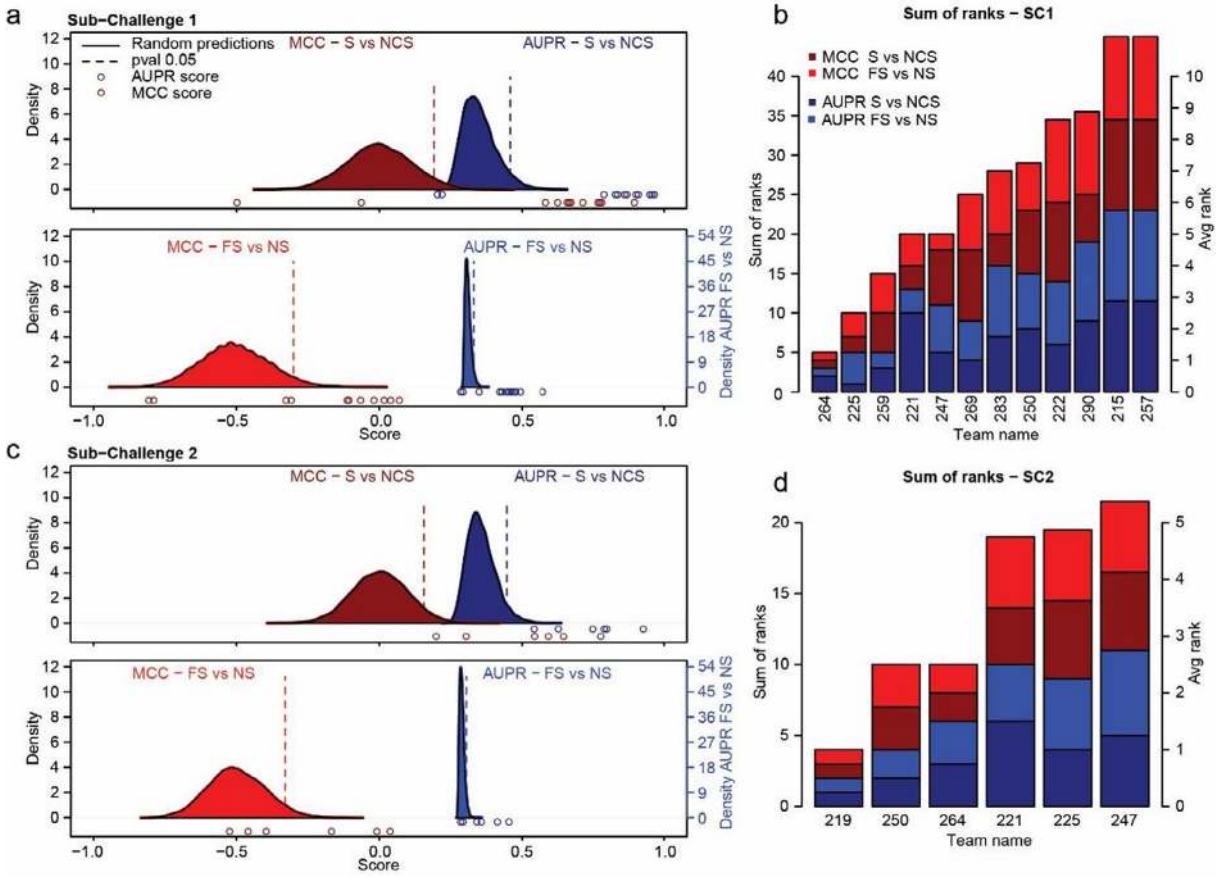

Fig. 2. Participants' prediction performances and final ranking. (a, c) Participants' scores (x-axis) relative to the null distribution (density curves) calculated from 10,000 random predictions. Dark blue and dark red (a-up, c-up) refer to the smoker (S) vs non-current smoker (NCS) task for area under precision recall (AUPR) and Matthew correlation coefficient (MCC), respectively. Blue and red (a-down, c-down) refer to former smoker (FS) vs never smoker (NS) task for AUPR and MCC, respectively. Blue and red circles identify participant's scores. Vertical dashed lines indicate the scores with P-values of 0.05 (smaller P-values are on the right of the dashed line). (b, d) Bar plots showing the sum of ranks (y-axis, left scale) and the average rank (y-axis, right scale) across all metrics and tasks for all teams for SC1 (b) and SC2 (d). A lower sum of rank implies better performance.

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

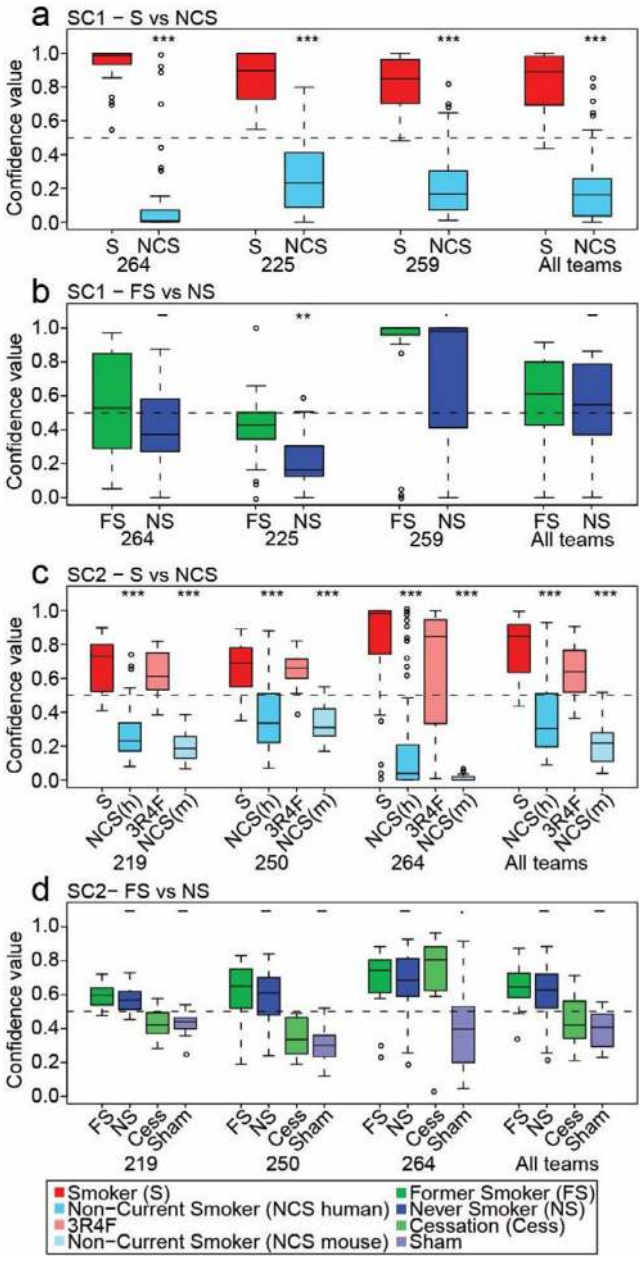

Fig. 3. Exposure class predictions by top performers and across all teams. Box plot showing the distributions of confidence scores (and median confidence scores for "All teams") for samples belonging to different exposure classes. The higher (close to 1) the value the higher the confidence that a subject is a smoker. Low values imply high confidence that the subject is a non-current smoker (NCS; i.e., former smoker (FS) or never smoker (NS)). (a) SC1: Smoker (S) vs NCS confidence score distributions for the three best-performing teams in SC1, and median confidence score distributions for all teams. (b) SC1: FS vs NS confidence score distributions for top three best-performing teams, and median confidence scores distribution of all teams. (c) SC2: S vs NCS (human) and 3R4F (exposed) vs NCS (non-exposed) (mouse) confidence score distributions for top three best-performing teams in SC2, and median confidence score distribution of all teams. (d) SC2: FS vs NS (human) and cessation (Cess) vs Sham (mouse) confidence score distributions for top three best-performing teams, and median confidence score distributions of all teams. [Wilcoxon-Mann-Whitney P-value, '.' < 0.1, '*' < 0.05, '**' < 0.01, '***' < 0.001, ('–' ≥ 0.1)].

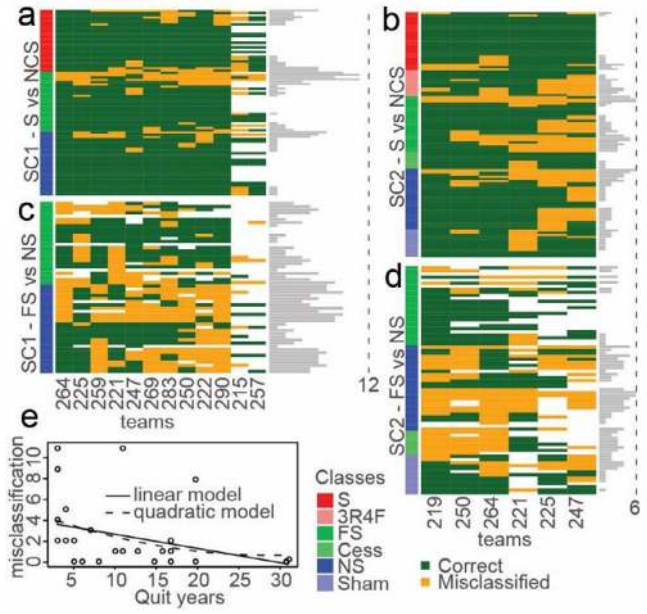

Fig. 4. Sample misclassifications. Sub-Challenge1 (SC1) (a, c) and SC2 (b, d) misclassifications shown as heatmaps. Teams are in columns arranged in decreasing order of performance from left to right. Subjects are in rows with the class label color as sidebar (smoker (S/3R4F), former smoker (FS/Cess), never smoker (NS/Sham)). Rows were clustered per class according to a binary distance between rows. Cells in green correspond to subjects correctly classified, cells in ochre correspond to misclassifications. White cells indicate the absence of a prediction. Horizontal bars show the number of subjects misclassified in each row. (e) Number of years since a FS quit smoking (x-axis) vs number of times the FS was misclassified. Linear and quadratic model fitting are reported.

SASH1, *AHRR*, and *LRRN3*, was identified by selecting genes with at least two co-occurrences across the signatures of the top six teams. The impact of gene signature size and co-linearity level on classification performance was investigated. The analysis was conducted using fivefold cross-validated training (with 10 repeats) and test datasets from SC1, separately. The most widely applied machine learning (ML) methods in the challenge were Random Forest (RF), support vector machine with linear kernel (svmLinear), partial least squares discriminant analysis (PLS), naive Bayes (NB), k-Nearest Neighbor (kNN), linear discriminant analysis (LDA), and logistic regression (LR). All possible combinations of the 18 genes of length 2–18 (i.e. 262,125 gene sets) were generated. Applying each of the seven ML methods to each gene set led to a total of 1,834,875 tested classification strategies. The level of co-linearity of genes within a gene set was reflected as the percentage of variance of the first principal component of the expression matrix restricted to that gene set. The performance of the 1,834,875 gene set-ML predictions (called "Top") was evaluated by computing MCC and AUPR scores. The performance of these "Top" gene sets were compared with that of gene sets (2–18 genes) randomly selected among the DEGs or all genes represented on the HG-U133_Plus_2 chip. The sampling process was repeated 1000 times for each gene set size, leading to a total of 17,000 random "DEG" or "All genes" gene sets. All ML methods applied for classification using the default parameters are available in the Caret R package version 6.0–41 (https://cran.r-project.org/web/packages/caret/index.html).

Impact of gene signature length on performances based on an ensemble learning method

A separate analysis was performed to measure the impact of varying gene signature lengths on classification performances. The Weka tool [46] was used to rank [47] and order attributes (genes) by weights of a Support Vector Machine (SVM) classifier. A list of the first 100 top-ranked genes is reported in Supplementary table 9 [48]. Performances were measured on a 10-fold stratified cross-validation on the training set of ML methods from the

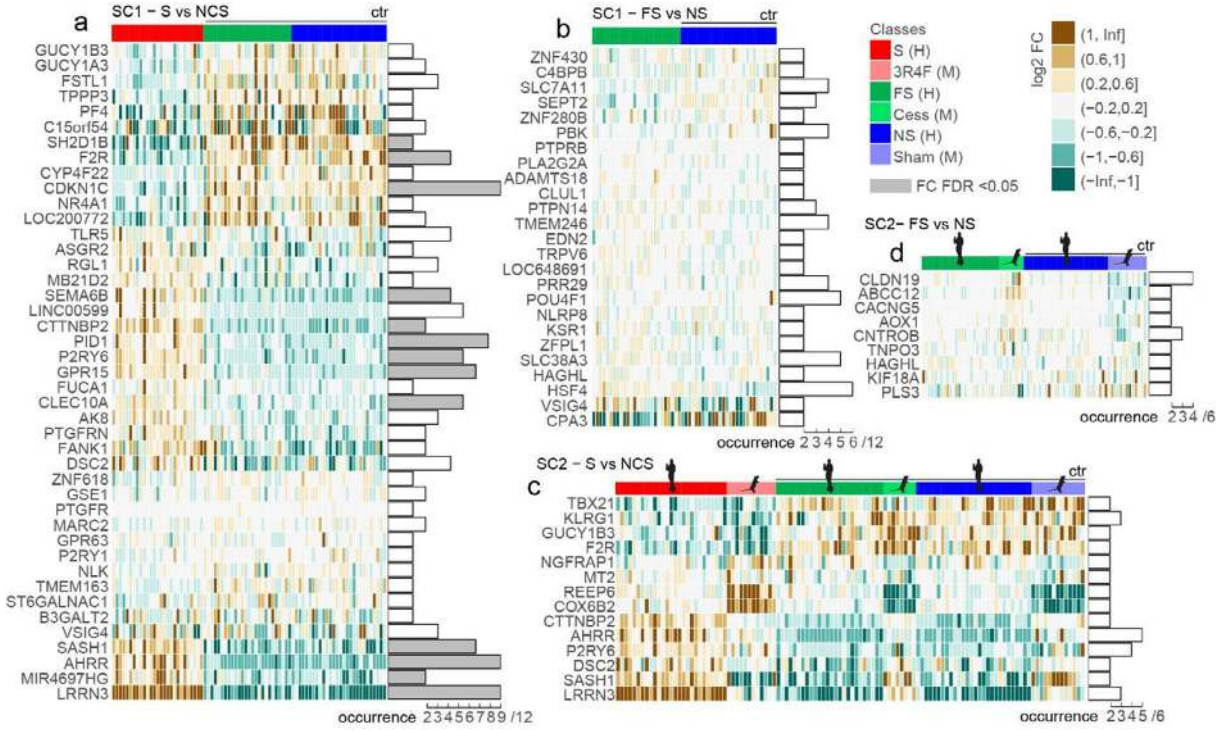

Fig. 5. Expression fold changes in the test dataset and co-occurrences of genes from consensus smoking exposure and cessation signatures. Differential gene expression heatmaps for the test datasets for (a, b) SC1 and (c, d) SC2. Subjects are in columns and grouped per class. Smokers (S) are in red (3R4F in light red), former smokers (FS) are in green (cessation (Cess) in light green), and never smokers (NS) are in blue (Sham in light blue). The respective control groups are annotated as ctr. (a) SC1: S vs NCS (ctr: FS + NS). (c) SC2: S vs NCS and 3R4F vs NCS (ctr: Cess + Sham). (b) SC1: FS vs NS. (d) SC2: FS vs NS and Cess vs Sham. Lengths of horizontal bars are proportional to the number of times a gene is selected as part of a signature. Gray bars denote genes for which the fold change (FC) is statistically significant (FDR < 0.05).

Python Scikit Learn library (http://scikit-learn.org). All classifiers were run with default parameter settings in all experiments. Performances were measured with AUPR and MCC metrics. The list of the 11 ML methods tested, including methods used by the best-performing teams, were: Stochastic Gradient Descent; Random Forest (RF); Extremely Randomized Trees; AdaBoost; Gradient Tree Boosting (GTB); Gaussian Naive Bayes; k-Nearest Neighbors (kNN); MultiLayer Perceptron (MLP); Linear Discriminant Analysis (LDA); Logistic Regression (LR); and Support Vector Classifier (SVC). The four best algorithms were selected for the ensemble learning method. The four ML methods were combined by an ensemble logic based on a soft voting mechanism: class label prediction was based on the argmax of the sums of the predicted probabilities of each ML method. The 4-classifier ensemble learning methods were trained on a dataset containing only expression values of the smallest gene signature (60 genes) by which all tested ML methods performed at 100% of accuracy during cross-validation.

Results

The Systems Toxicology challenge addressed the problem of identification of blood exposure response markers, in the form of gene expression-based signatures, predictive of smoking exposure and cessation status in human (SC1) and across both human and rodent species (SC2) for translational toxicology (Fig.1a). After training blood-based signature classification models, participants applied their models on independent unlabeled gene expression data for class prediction. Participants' eligibility for scoring was conditional to their compliance with the challenge rules (1, submission completeness; 2, predictions using inductive classification models; and 3, gene signature length ≤40 genes).

Participation, scores, and final ranking

The Systems Toxicology challenge attracted worldwide participation (Supplementary Fig. 2). Among the 61 teams registered (165 participants; 47% China, 21% Europe, 10% USA, and 9% India), 23 and 15 teams provided submissions for SC1 and SC2, respectively. Among them, 12 and six teams complied with all challenge rules, and therefore were eligible for scoring. MCC and AUPR scores computed on anonymized teams' predictions were compared with an empirical null distribution generated from 10,000 random prediction sets and were considered as significant when they exceeded the 95th percentile of the null distribution (Fig. 2a and c, and Supplementary table 4).

MCC scores obtained from random predictions for task 2, former smokers (FS) vs never smokers (NS), were centered on −0.5 because all samples that were misclassified in task 1 were considered as incorrect predictions for scoring task 2 predictions ("penalization"). For both sub-challenges, 80–100% and 50–80% of the teams obtained significant MCC and AUPR scores for smoker (S) vs non-current smoker (NCS; i.e., FS and NS) and FS vs NS classifications, respectively (Fig. 2a and c). Final ranking was based on averaging the score-based ranks across metrics and tasks for each sub-challenge (Fig. 2b and d, and Supplementary table 4). Teams 264 and 219 were declared best performers for SC1 and SC2, respectively. Teams 225 and 259 were ranked second and third for SC1, respectively, and teams 264 and 250 were ranked second equal for SC2.

To understand the impact of the penalization scheme mentioned above on score significance for the discrimination of FS vs NS, the top teams of each sub-challenge (except team 225) provided us with predictions for all samples (as opposed to samples predicted only as NCS) using the same classification models. The results showed (Supplementary Fig. 3) that, for SC1, only team

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

264's predictions reached significance (P-value < 0.05) for both the MCC and AUPR scores and, for SC2, the predictions of teams 250 and 264 were significant only for the AUPR score or MCC score, respectively. These observations indicated that, overall, the penalization procedure led to team scores that were higher relative to the null distribution of scores than they should have been when all class samples were predicted. Moreover, the shift of best-performing teams toward lower or non-significant scores highlighted the difficulty to discriminate FS from NS using only gene expression data from blood.

Prediction of smoking exposure status was possible in human and across species using blood gene expression, while prediction of cessation status was more challenging

As classification predictions, participants were requested to submit confidence values P (between 0 and 1) and 1-P that a sample belonged to class/group 1 (e.g., S) and class/group 2 (e.g., NCS), respectively. Confidence value distributions per actual group/class are displayed as boxplots in Fig. 3. For both sub-challenges, significant differences between the smoke-exposed and non-current smoke-exposed groups were calculated based on log odds transformed confidence values. A clear separation between both groups (S/3R4F vs NCS) was observed in boxplots generated with individual and aggregated confidence values of the top three teams and all teams, respectively (Fig. 3a and c). By contrast, the discrimination between FS and NS subjects was difficult, as illustrated by the lack of significant difference in confidence value distributions (Fig. 3b and d). Despite this observation, the medians of confidence values for FS/cessation groups tended to be systematically higher than the medians for NS/Sham groups for individual and aggregated results (Supplementary Fig. 4). Overall, these results show that blood gene expression data were informative to build human and species-independent blood-based gene signature classification models for smoking exposure status, but weakly informative to derive signature classification models predictive of cessation status.

Samples from formerly smoke-exposed group were more frequently misclassified in both sub-challenges

Individual sample classification was closely investigated to identify whether some samples were systematically misclassified across teams. In general, most of the subjects were correctly classified as S vs NCS in both sub-challenges, as shown by the majority of green color cells in the heatmaps (Fig. 4a and b) representative of correctly classified samples.

Misclassification increased significantly for former smokers as compared to both never (P-value = 5.58e-06) and current smokers (P-value = 0.0032). Finally, the p-value for the misclassification for former smokers compared to any other group (current and never smokers) was lower than 0.001 (Table 1).

Samples from FS were misclassified by at least 50% and 70% of teams in SC1, 45% and 75% (vs 32% of the total number of all human test samples), respectively. For SC2, the percentages of FS samples misclassified by at least 50% and 70% of teams were 48% and 67% (v.s. 30% of the total number of all human and mouse test samples), respectively. In both SC1 and SC2, the classification of samples as FS and NS resulted in higher misclassification rate compared with their classification as S vs NCS, as reflected by the similar proportion of green and ochre colored heatmap cells corresponding to correctly classified and misclassified samples, respectively (Fig. 4c and d). Overall, the results show that samples from the FS group were the most frequently misclassified (for human).

The wisdom of crowds enabled the identification of consensus blood-based smoking exposure and cessation gene signatures

Consensus smoking exposure and cessation signatures were obtained by considering genes present in the signatures of at least two teams (Fig. 5; Supplementary table 5). The consensus human and species-independent gene signatures that could discriminate smoking exposure status contained 43 and 14 genes, respectively, and those identified as predictors of the cessation status included 25 and nine genes, respectively. Human and species-independent smoking exposure signatures had eight genes in common, while no gene overlap was observed for human and species-independent cessation signatures. In general, the expression fold change of genes in the smoking exposure signatures comparing S/3R4F vs NCS were higher than the expression fold change of genes in the cessation signatures comparing FS/cessation vs NS/Sham. In addition, the genes in the smoking exposure signatures showed clear expression patterns separating smoke-exposed and non-current smoke-exposed subjects (Fig.5a and c).

Not surprisingly, this separation was even more pronounced on the heatmap generated with the training dataset (Supplementary Fig. 5). For the cessation signatures, gene expression patterns were observed for the training dataset (Supplementary Fig. 5); however, the patterns became barely visible for the test datasets (Fig.5b and d), indicating that gene expression level differences were subtler between the FS/cessation and NS/Sham groups, and that the selected genes only weakly generalized as discriminative features across independent datasets. Notably, some genes in the smoking exposure (e.g., *LRRN3*, *MT2*, *P2RY6*) and cessation (e.g., *AOX1*, *ABCC12*) signatures showed opposite fold change directions in human and mouse for the same exposure groups such as S and 3R4F or FS and cessation (Fig.5c and d).

Table 1
Sample misclassification across teams. The number (and percentage) of misclassified samples per sub-challenge (SC) and task are shown. For SC1 (top) and SC2 (bottom), the total number of samples (including group and sex) misclassified by at least 50% and 70% of the teams are reported. M: Male; F: Female; FS: Former Smoker; NS: Never Smoker; S: Smoker.

Sub-challenge	% Teams	Total misclassified samples	Misclassified samples per class	Total samples per class
SC1	≥50%	11 (4 M/7 F)	5 (45%) FS 3 (27%) NS 3 (27%) S	26 FS 28 NS 27 S
	≥70%	4 (2 M/2 F)	3 (75%) FS 1 (25%) NS 0 (0%) S	
SC2	≥50%	25 (5 M/20 F)	12 (48%) FS 6 NS + 1 Sham (28%) 2 S + 4 3R4F (24%)	26 FS/8 Cess 28 NS/13 Sham 27 S/12 3R4F
	≥70%	6 (2 M/5 F)	4 (67%) FS 1 (17%) NS 1 (17%) S	

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

Genes coding for cell surface receptors and molecules involved in proximal pathway signaling were over-represented in consensus blood smoking exposure signatures

For biological interpretation of the blood-based gene signatures, a pathway/process over-representation analysis was performed using Broad MSigDB, DAVID and Ingenuity Pathway Analysis. Only the human smoking exposure 'consensus' signature that counted 43 genes showed significant pathway/process over-representation. Lack of significant pathway over-representation for the other consensus signatures may be due to their lower size, and hence reduced power. For the human smoking gene signature, the over-represented MSigDB gene sets (FDR $\leq$ 0.05) were associated with cell surface receptors such as G-protein coupled receptors (GPCR) and purinergic receptors, and intracellular signaling such as G-proteins (Table 2).

Genes that had transcription factor binding site for *NFAT*, *FOXO4* or *BACH2* in their promoter sequences were also over-represented (Table 2). Similar results were identified with DAVID's functional annotation clustering method, which groups category terms corresponding to pathway/process as well as protein sequence features and functions into annotation clusters (Supplementary table 6). The terms associated with the most significant cluster were related to protein localization corresponding to transmembrane, integral to membrane, membrane and also to glycoprotein, glycosylation sites. The second and third clusters included terms related to coagulation, wound healing and nucleotide receptor activity, purinergic nucleotide receptor activity, cell surface receptor signal transduction, GPCR protein signaling, phospholipase C activity. Other clusters included terms such as inflammation, defense response, blood circulation, immunoglobulin, phosphorylation, kinase activity, ion binding. In the "Diseases and Functions" category, Ingenuity Pathway Analysis highlighted top significantly over-represented processes associated with carbohydrate metabolism, cardiovascular disease and function (e.g., endothelial permeability, hypertension, myocardial infarction), hematological disease and function (e.g., blood coagulation, thrombosis) (data not shown). In addition, the literature was queried to find whether genes in the consensus human smoking exposure signature had previously been associated with smoking in blood. Interestingly, and despite the fact that the feature selection was not based on prior knowledge, genes showing the highest co-occurrences across teams were, in general, the most reported in the literature to be associated with smoking (Supplementary table 7).

Gene set combinations from an 18 gene-based consensus signature from the top six teams were informative and outperformed "DEGs" and "All Genes"-derived gene sets for smoking exposure status class prediction

The impact of gene signature size and co-linearity level on the performance of smoking exposure status class prediction was explored using the 18 gene-based consensus signature from the top six teams' predictions. MCC and AUPR scores were calculated to evaluate the performance of all possible combinations of signatures of lengths 2–18 with ML-based class predictions (Fig. 6 and Supplementary Fig. 6).

The prediction performance increased with gene set size and gradually stabilized with longer sets, including up to 18 genes in both training (cross-validation, CV) (for CV, MCC = 0.57 for size = 2 and MCC = 0.91 for size = 18) and test sets (for test, MCC = 0.42 for size = 2 and MCC = 0.77 for size = 18) (Fig. 6a). Prediction performances reached maximum when the co-linearity level (reflected by the percentage of variance represented by the first principal component computed from the gene set expression matrix) of genes in the "Top" gene sets ranged between 50% and 60%, and then decreased with increased co-linearity (Fig. 6b). Considering that the "Top" gene sets were composed of the signature genes from different teams and were already quite diverse, combining genes that are to some extent co-linear may strengthen the prediction. Performances decreased with increased co-linearity of genes within gene sets from DEGs (Fig. 6b). In general, gene sets from "Top", "DEG", and "All Genes" gave the best, middle, and worst performances, respectively (Fig. 6a, b and c). In addition, performances derived from CV outperformed those computed for the test set (Fig. 6a, b and c). Performance metrics obtained with various ML methods showed similar patterns (Fig. 6b), and therefore, were aggregated to facilitate the visualization of results (Fig. 6a and c). Overall, the results indicated that blood genes from the 18 gene-based consensus signature were informative and had high predictive power for smoking exposure status when combined.

Table 2
Over-representation analysis of biological pathways/processes associated with the consensus human smoking exposure gene signature. List of pathways, biological processes, and transcription factor targets enriched in the consensus gene signature. Statistical significance is reported in the second column.

MSigDB Canonical Pathways (Canonical pathways, BioCarta, KEGG, Reactome) Gene Sets	$-LOG_{10}$(FDR)
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	2.05
REACTOME_G_ALPHA_Q_SIGNALLING_EVENTS	1.99
REACTOME_GPCR_LIGAND_BINDING	1.99
REACTOME_GASTRIN_CREB_SIGNALLING_PATHWAY_VIA_PKC_AND_MAPK	1.99
REACTOME_P2Y_RECEPTORS	1.94
REACTOME_HEMOSTASIS	1.94
REACTOME_NUCLEOTIDE_LIKE_PURINERGIC_RECEPTORS	1.81
KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	1.81
REACTOME_NITRIC_OXIDE_STIMULATES_GUANYLATE_CYCLASE	1.49
MSigDB - GO Biological Process	
SIGNAL_TRANSDUCTION	5.14
CELL_SURFACE_RECEPTOR_LINKED_SIGNAL_TRANSDUCTION_GO_0007166	4.53
INTRACELLULAR_SIGNALING_CASCADE	2.20
G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	1.34
MSigDB - Transcription Factor Targets (Top3)	
TGGAAA_V$NFAT_Q4_01	2.63
TTGTTT_V$FOXO4_01	2.61
V$BACH2_01	1.73

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

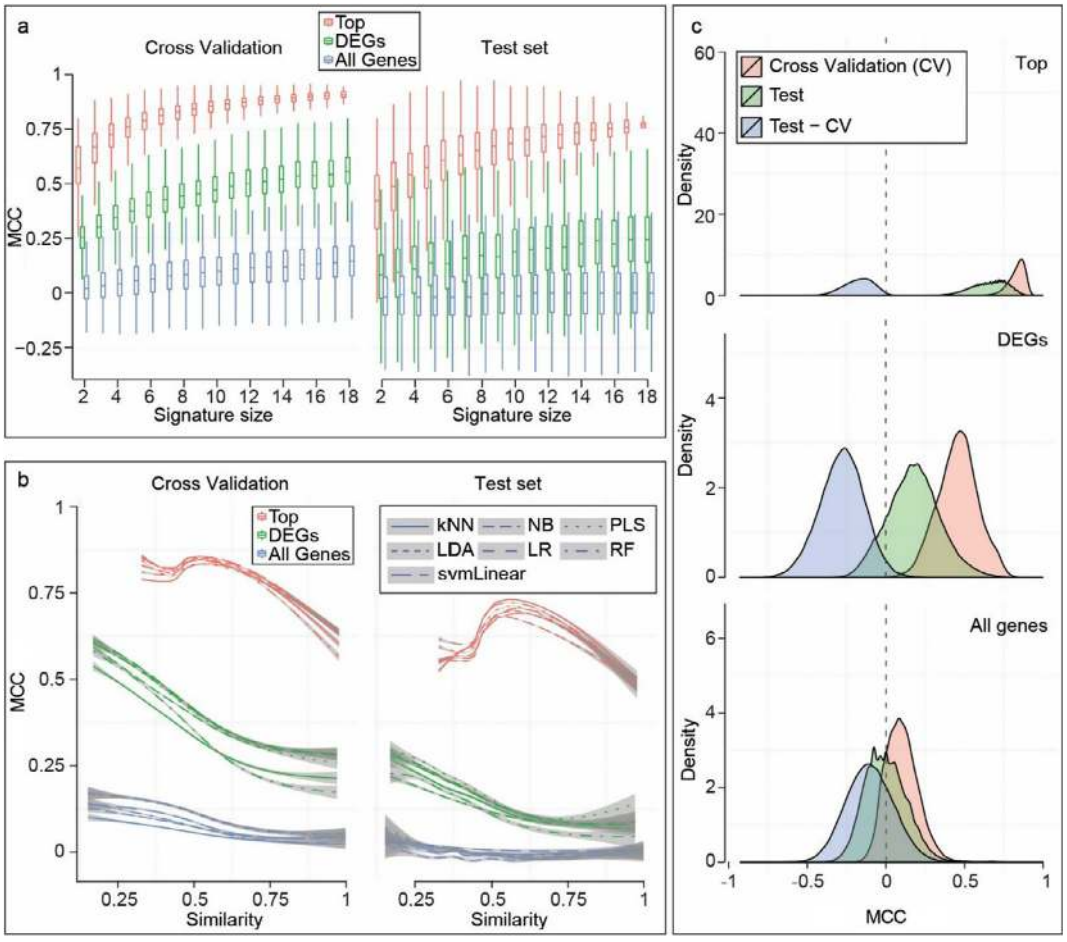

Fig. 6. Performance versus signature size and gene similarity. (a) Matthew correlation coefficient (MCC) score versus gene signature size for cross-validation and test dataset. Features were selected from the list of (i) "Top" genes (red), i.e., genes selected frequently by participants as part of the signature; (ii) "DEGs" (green), list of differentially expressed genes; (iii) "All Genes" (light blue), all measured genes. (b) MCC performance versus coefficient of similarity between genes in the signature. Seven different machine learning classifier were tested: (Random Forest (RF), support vector machine with linear kernel (svmLinear), partial least squares discriminant analysis (PLS), naive Bayes (NB), k-Nearest Neighbor (kNN), linear discriminant analysis (LDA), and logistic regression (LR)). (c) Distributions of MCC scores in CV (red) and test set (green) data, plus distribution of the differences (light blue), for "Top" (top), "DEGs" (middle), and "All genes" (bottom) selections.

Performances degraded for long gene signatures (>40 genes)

A separate analysis was conducted to understand the impact of the length of gene signatures selected as discriminative features on classification performances. Genes were selected by SVM, and ranked according to the feature selection SVM (see Material and Methods for details). The cross-validation performances of the different methods are reported in Supplementary Fig. 7 and Supplementary table 8. Ensemble learners based on decision trees methods (AdaBoost, GTB, RF, and ERT) had accuracies in the range 90–96%, but did show significant improvements with increased signature lengths. RF, ERT, and GTB in particular, showed oscillating trends in accuracy. The kNN and Gaussian Naive Bayes methods had accuracies in the ranges 96–99% and 94–96%, respectively. The Stochastic Gradient Descent method had high accuracies of around 99–100%, with peaks limited to some signature size ranges. Four methods (MLP, LR, LDA, and SVC) were robust enough to maintain 100% accuracies throughout all signature lengths larger than 30 genes and were selected for the ensemble learner. Ensemble learner performances on the test set are reported in Fig. 7 (and Supplementary fig. 8).

As the graph shows, the best prediction of the ensemble learner on the S vs NCS task was obtained when a signature length of 40 genes was selected (accuracy = 92.6%, AUPR = 95.8%, and MCC = 83.2%). It is worth noting that by further increasing/decreasing the signature length, the prediction accuracy (as well as the AUPR and MCC scores) decreased. A signature length of 40 genes was found to be optimal on the test set even though, from the cross-validation analysis, 40 genes were not sufficient for the ensemble learner to reach top performances.

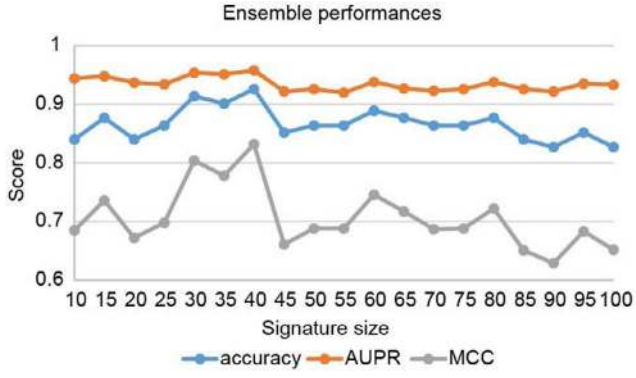

Fig. 7. Score versus signature size (≥10 genes). Performances of the ensemble learner (top four methods from cross-validation). Performance accuracy (blue), Area under the precision-recall (AUPR) curve (orange), and Matthew correlation coefficient (MCC) (gray) scores degraded when gene signature length increased above 40 genes.

Discussion

Crowdsourcing is a powerful approach to solve scientific problems, and also to independently verify methods, results, and conclusions [31]. Systems toxicology is an emerging discipline that requires the development of advanced computational methods for the analysis of large-scale datasets and the extrapolation of predictive toxicological outcomes and risk estimates. Here, we summarize the results and lessons from the sbv IMPROVER Systems Toxicology computational challenge that was opened to the scientific community (https://sbvimprover.com).

The predictions submitted by the participants in the challenge demonstrated that blood gene expression data were informative enough to identify specific markers and train models predictive of smoking exposure status in human and across human and mouse species. This finding confirmed that smoke has a prominent effect that propagates in blood cells distal to lungs, the primary site of exposure [20,27]. Best-performing teams succeeded in (i) developing inductive models able to predict the class of new independent individual samples with high levels of performance (AUPR $\geq$ 0.93 for SC1 and SC2), and (ii) extracting parsimonious signatures that did not exceed 40 genes. These results have implications for the development of blood-based molecular-diagnostics tool for monitoring an exposure response to cigarette smoking. These results may also be useful for assessing the exposure response to the new generation of non-combustible tobacco products with the potential to reduce individual risk. Applying a human blood-based smoking exposure response gene signature previously identified [27], Martin et al. demonstrated that samples from conventional cigarette smokers who switched to a THS2.2 shifted away from the S group and classified closer to the NCS group after only 5 days of switching [49]. These results were independently confirmed by the crowd in the context of this challenge [31].

Although different computational approaches were used for the challenge, the wisdom of crowds enabled the identification of robust consensus blood-based human and species-independent smoking exposure signatures, including genes with high co-occurrences across qualified teams, namely *AHRR*, *CDKN1C*, *LRRN3*, *PID1*, *GPR15*, *SASH1*, *CLEC10A*, *LINC00599*, and *P2RY6*. These genes overlapped remarkably with genes present in previously published blood-based smoking gene signatures derived from the same datasets, constituting an independent confirmation of a signature proposed by Martin et al. [27] and signatures from a meta-analysis of six studies that included a total of 10,233 subjects [50]. Many genes of the consensus signature are reported to undergo differential regulation at the mRNA and/or DNA methylation level following exposure to cigarette smoke (references in Supplementary table 7). Epigenetic-dependent/independent transcriptional regulation and/or blood cell population enrichment modifications may explain smoke-induced molecular changes observed in whole blood [51,52]. Further blood cell type-specific investigations may provide a deeper understanding of the contribution of each subpopulation to the smoking signature. For example, among the genes with the strongest association with smoking, *P2RY6*, *PID1*, and *SASH1* were identified in the transcriptome of monocytes, and *LRRN3* may be expressed specifically in T lymphocyte subpopulation [53,54]. Regarding the consensus species-independent smoking exposure signature, species-specific regulation mechanisms and/or blood cell population composition may account for the opposite directionality of expression changes observed in human and mouse for a subset of the genes.

Most genes present in the consensus smoking signature encode cell surface receptors and molecules involved in glycoprotein and innate immune pattern recognition/binding (*ASGR2*, *TLR5*, *P2RY6*, *P2RY1*), immune regulation (*VSIG4*), coagulation (*F2R*, *PF4*), and signal transduction such as G-proteins (*GPR15*, *GPR63*), guanylate cyclase (*GUCY1A3*, *GUCY1B3*), and kinases (*AK8*, *NLK*). These functional characteristics suggest that exposure to smoke modifies the properties of circulating blood cells seen as "sentinels" for sensing and reacting to their environment. For example, purinergic receptors present in many cell types recognize extracellular nucleotides (ATP, ADP, UDP) that are actively released or passively leaked from damaged or dying cells, as danger signals and trigger inflammation. Interestingly, it has been reported that nucleotide levels are elevated in plasma of active smokers with peripheral artery diseases [55]. The purinergic receptors may also be linked to the development of diseases as their inhibition can reduce/prevent the development of atherosclerosis and smoke-induced lung injury and emphysema [56,57]. Notably, P2Y6 receptor expression was found to be one of the most strongly induced genes in lung tissue of mice exposed to smoke [56]. Other genes present in the consensus signature such as *SASH1* may link smoking and atherosclerosis [58]. Therefore, the consensus blood smoking signature may not only include genes that reflect a response to smoke exposure, but also genes that may play roles in smoke-related disease pathogenesis in the long term. This hypothesis was supported by Huan et al. who found significant associations between their whole blood-based cigarette smoking signature and human complex diseases and traits [50].

Unlike smoking status, predicting cessation status that differentiates FS and NS individuals remains challenging, as reflected by the lower prediction performances and gene signature overlap across teams. This outcome indicates that blood gene expression data alone may not be sufficiently informative to discriminate FS and NS individuals. However, the observations that blood samples from FS misclassification increased significantly as compared to both never and current smokers suggested larger heterogeneity of the gene expression profiles compared with S and NS. Either FS were not fully compliant with quitting smoking or smoke-induced molecular changes remain persistent in blood even after periods of cessation. Smoking and cessation histories vary from one subject to another and related parameters (i.e., intensity, duration) and/or other data modalities (e.g., DNA methylation) may also need to be included in the modeling process. Zhang et al. [59] reported a linear relationship between cotinine concentration and DNA methylation levels at site cg05575921 (*AHRR*). Another study found that DNA methylation at *F2RL3* was dependent on the interaction of pack years and time since quit [60]. Gene expression and DNA methylation levels in FS have been shown to reverse with time [60,61] and approach those of NS several years after cessation [62], although persistent epigenetic markers of smoking have been detected decades (>35 years) after cessation [61]. In line with this report, we observed that FS who most recently quit smoking were more frequently misclassified as those who ceased smoking for longer time periods (Fig.4e).

Genes identified to be consensus genes were not necessarily differentially expressed between classes. The identification of a highly predictive gene signature does not rely on selecting only DEGs (although informative to some extent), but rather on building the right combination of genes that, as a set, have high discrimination power between classes. We have found that the magnitude and robustness of classification performance were dependent on finding the right balance of the number and co-linearity of genes (50–60% here; same gene expression patterns) to build the signature. Indeed, irrespective of the methodology applied for classification, performances started to degrade when the list of genes became too short or increased beyond a certain size (>40 genes). These observations highlighted the importance of a feature selection step. Once the most relevant features were identified, the ML methods used for classification did not necessarily show large

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004

differences in the performance accuracies of the sample classification. Another important learning was that cross-validation, irrespective of the way a signature was built, overestimated classification performances, which may lead to over-fitting [11]. Therefore, validation of a signature in independent cohorts is critical to ensure robustness and generalization of a predictive model.

In conclusion, the Systems Toxicology challenge outcomes showed that blood gene expression data were sufficiently informative to predict exposure to smoke in human and across species, but may not be sufficient for cessation status prediction. The crowd succeeded in developing robust inductive predictive models and identifying concise human and species-specific signatures that included genes with large consensus across teams. Post-challenge computational analysis also highlighted the importance of the feature selection step in the process of building a classifier and the need for validation of a gene signature in independent cohorts. Overall, leveraging the power and wisdom of the crowd in this challenge demonstrated the importance of independent and unbiased evaluations of data and computational methods to provide learnings, confirmation, and confidence in scientific conclusions in systems toxicology.

Acknowledgements

We thank all participants for their active participation to the sbv IMPROVER Systems Toxicology computational challenge; Anouk Ertan, Laure Cannesson, and David Page for their support on project management and communication; Prof. Leonidas Alexopoulos, Dr. Alberto de la Fuente, and Prof. Rudiyanto Gunawan as part of the external Scoring Review Panel for their expert support on the scoring procedure; Dr. Bjorn Titz for developing the ortholog mapping functionalities; Dr. Gregory Vuillaume and Dr Athanasios Kondylis for some support in statistical analysis; Filipe Bonjour and Sylvain Gubian for IT support; Dr. Alain Sewer for his scientific expertise and discussions during the challenge preparation. The work was fully funded by Philip Morris Products S.A.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.comtox.2017.07.004.

References

[1] H.G. LaBreche et al., Peripheral blood signatures of lead exposure, PLoS One 6 (8) (2011) e23043, http://dx.doi.org/10.1371/journal.pone.0023043.
[2] P.R. Bushel et al., Blood gene expression profiling of an early acetaminophen response, Pharmacogenomics J. (2016), http://dx.doi.org/10.1038/tpj.2016.8.
[3] P. Joseph, C. Umbright, R. Sellamuthu, Blood transcriptomics: applications in toxicology, J. Appl. Toxicol. 33 (11) (2013) 1193–1202, http://dx.doi.org/10.1002/jat.2861.
[4] R.S. Thomas et al., Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework, Toxicol. Sci. 136 (1) (2013) 4–18, http://dx.doi.org/10.1093/toxsci/kft178.
[5] P. Farmer et al., A stroma-related gene signature predicts resistance to neoadjuvant chemotherapy in breast cancer, Nat. Med. 15 (1) (2009) 68–74, http://dx.doi.org/10.1038/nm.1908.
[6] G.J. Gordon et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Res. 62 (17) (2002) 4963–4967.
[7] P. Wirapati et al., Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures, Breast Cancer Res. 10 (4) (2008) R65, http://dx.doi.org/10.1186/bcr2124.
[8] J.D. Zhang et al., Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity, Pharmacogenomics J. 14 (3) (2014) 208–216, http://dx.doi.org/10.1038/tpj.2013.39.
[9] C.O.E. Team, Blood-based gene expression analysis platform for molecular diagnostics, Clinical OMICS 1 (14) (2014).
[10] Z. Zhang, An In Vitro Diagnostic Multivariate Index Assay (IVDMIA) for ovarian cancer: harvesting the power of multiple biomarkers, Rev. Obstet. Gynecol. 5 (1) (2012) 35–41.
[11] S. Barbash, H. Soreq, Statistically invalid classification of high throughput gene expression data, Sci. Rep. 3 (2013) 1102, http://dx.doi.org/10.1038/srep01102.
[12] V.N. Vapnik, Statistical learning theory, Wiley, New York, 1998.
[13] P. Meyer et al., Industrial methodology for process verification in research (IMPROVER): toward systems biology verification, Bioinformatics 28 (9) (2012) 1193–1201, http://dx.doi.org/10.1093/bioinformatics/bts116.
[14] E. Bilal et al., A crowd-sourcing approach for the construction of species-specific cell signaling networks, Bioinformatics 31 (4) (2015) 484–491, http://dx.doi.org/10.1093/bioinformatics/btu659.
[15] C. Poussin et al., The species translation challenge-a systems biology perspective on human and rat bronchial epithelial cells, Sci. Data 1 (2014) 140009, http://dx.doi.org/10.1038/sdata.2014.9.
[16] K. Rhrissorrakrai et al., Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge, Bioinformatics 31 (4) (2015) 471–483, http://dx.doi.org/10.1093/bioinformatics/btu611.
[17] sbv, I.p.t., et al., Community-reviewed biological network models for toxicology and drug discovery applications. Gene Regul. Syst. Biol., 2016. 10: p. 51–66. doi: 10.4137/GRSB.S39076.
[18] A.L. Tarca et al., Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge, Bioinformatics 29 (22) (2013) 2892–2899, http://dx.doi.org/10.1093/bioinformatics/btt492.
[19] B. Messner, D. Bernhard, Smoking and cardiovascular disease: mechanisms of endothelial dysfunction and early atherogenesis, Arterioscler. Thromb. Vasc. Biol. 34 (3) (2014) 509–515, http://dx.doi.org/10.1161/ATVBAHA.113.300156.
[20] R. Faner et al., Systemic inflammatory response to smoking in chronic obstructive pulmonary disease: evidence of a gender effect, PLoS One 9 (5) (2014) e97491, http://dx.doi.org/10.1371/journal.pone.0097491.
[21] H.K. Na et al., Tobacco smoking-response genes in blood and buccal cells, Toxicol. Lett. 232 (2) (2015) 429–437, http://dx.doi.org/10.1016/j.toxlet.2014.10.005.
[22] S. Boue et al., Cigarette smoke induces molecular responses in respiratory tissues of ApoE(-/-) mice that are progressively deactivated upon cessation, Toxicology 314 (1) (2013) 112–124, http://dx.doi.org/10.1016/j.tox.2013.09.013.
[23] B. Halvorsen et al., Effect of smoking cessation on markers of inflammation and endothelial cell activation among individuals with high risk for cardiovascular disease, Scand. J. Clin. Lab. Invest. 67 (6) (2007) 604–611, http://dx.doi.org/10.1080/00365510701283878.
[24] B. Phillips et al., A 7-month cigarette smoke inhalation study in C57BL/6 mice demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product, Food Chem. Toxicol. 80 (2015) 328–345, http://dx.doi.org/10.1016/j.fct.2015.03.009.
[25] P. Beineke et al., A whole blood gene expression-based signature for smoking status, BMC Med. Genomics 5 (2012) 58, http://dx.doi.org/10.1186/1755-8794-5-58.
[26] R. Joehanes et al., Epigenetic signatures of cigarette smoking, Circ Cardiovasc. Genet. 9 (5) (2016) 436–447, http://dx.doi.org/10.1161/CIRCGENETICS.116.001506.
[27] F. Martin et al., Identification of gene expression signature for cigarette smoke exposure response–from man to mouse, Hum. Exp. Toxicol. 34 (12) (2015) 1200–1211, http://dx.doi.org/10.1177/0960327115600364.
[28] B. Titz et al., Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects, J Proteomics 128 (2015) 306–320, http://dx.doi.org/10.1016/j.jprot.2015.08.009.
[29] C. Haziza et al., Assessment of the reduction in levels of exposure to harmful and potentially harmful constituents in Japanese subjects using a novel tobacco heating system compared with conventional cigarettes and smoking abstinence: a randomized controlled study in confinement, Regul. Toxicol. Pharmacol. 81 (2016) 489–499, http://dx.doi.org/10.1016/j.yrtph.2016.09.014.
[30] C. Haziza et al., Evaluation of the Tobacco Heating System 2.2. Part 8: 5-Day randomized reduced exposure clinical study in Poland, Regul. Toxicol. Pharmacol. 81 (Suppl 2) (2016) S139–S150, http://dx.doi.org/10.1016/j.yrtph.2016.11.003.
[31] C. Poussin et al., 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. (2017), http://dx.doi.org/10.1021/acs.chemrestox.6b00345.
[32] B. Phillips et al., An 8-month systems toxicology inhalation/cessation study in Apoe-/- mice to investigate cardiovascular and respiratory exposure effects of a candidate modified risk tobacco product, THS 2.2, compared with conventional cigarettes, Toxicol. Sci. 151 (2) (2016) 462–464, http://dx.doi.org/10.1093/toxsci/kfw062.
[33] M.R. Smith et al., Evaluation of the Tobacco Heating System 2.2. Part 1: Description of the system and the scientific assessment program, Regul. Toxicol. Pharmacol. 81 (Suppl 2) (2016) S17–S26, http://dx.doi.org/10.1016/j.yrtph.2016.07.006.
[34] Team, R.D.C., R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2008.

[35] M.N. McCall, B.M. Bolstad, R.A. Irizarry, Frozen robust multiarray analysis (fRMA), Biostatistics 11 (2) (2010) 242–253, http://dx.doi.org/10.1093/biostatistics/kxp059.
[36] Matthew N. McCall, R.A.I., mouse4302frmavecs: Vectors used by frma for microarrays of type mouse4302. R package version 1.3.0.
[37] Matthew N. McCall, R.A.I., hgu133plus2frmavecs: Vectors used by frma for microarrays of type hgu133plus2. R package version 1.3.0.
[38] F. Petrescu, S.C. Voican, I. Silosi, Tumor necrosis factor-alpha serum levels in healthy smokers and nonsmokers, Int. J. Chron. Obstruct. Pulmon. Dis. 5 (2010) 217–222.
[39] M.E. Ritchie et al., limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47, http://dx.doi.org/10.1093/nar/gkv007.
[40] T.A. Eyre et al., HCOP: a searchable database of human orthology predictions, Brief Bioinform. 8 (1) (2007) 2–5, http://dx.doi.org/10.1093/bib/bbl030.
[41] S.G. Isabelle Guyon, Masoud Nikravesh, Lotfi A. Zadeh, Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing). Springer.
[42] P. Baldi et al., Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (5) (2000) 412–424.
[43] G.K. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 2004. 3: p. Article3. doi: 10.2202/1544-6115.1027.
[44] A. Subramanian et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. U S A 102 (43) (2005) 15545–15550, http://dx.doi.org/10.1073/pnas.0506580102.
[45] G. Dennis Jr et al., DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol. 4 (5) (2003) P3.
[46] M.A.H. Eibe Frank, Ian H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques". Fourth Edition ed, ed. M. Kaufmann. 2016.
[47] M. Guarracino, S.C., D. Feminiano, G. Toraldo, P. Pardalos. Current classification algorithms for biomedical applications. in Centre de Recherches Mathmatiques CRM Proceedings & Lecture Notes of the American Mathematical Society. 2008.
[48] I. Guyon, J.W., S. Barnhill, V. Vapnik., Gene selection for cancer classification using support vector machines, in Machine Learning, 2002. p. 389-422.
[49] F. Martin et al., Evaluation of the tobacco heating system 2.2. Part 9: Application of systems pharmacology to identify exposure response markers in peripheral blood of smokers switching to THS2.2, Regul. Toxicol. Pharmacol. 81 (Suppl 2) (2016) S151–S157, http://dx.doi.org/10.1016/j.yrtph.2016.11.011.
[50] T. Huan et al., A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking, Hum. Mol. Genet. (2016), http://dx.doi.org/10.1093/hmg/ddw288.
[51] M. Bauer et al., Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation, Clin. Epigenetics 7 (2015) 83, http://dx.doi.org/10.1186/s13148-016-0249-7.
[52] M. Bauer et al., A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood, Clin. Epigenetics 7 (2015) 81, http://dx.doi.org/10.1186/s13148-015-0113-1.
[53] J.P. Chou et al., Accelerated aging in HIV/AIDS: novel biomarkers of senescent human CD8+T cells, PLoS One 8 (5) (2013) e64702, http://dx.doi.org/10.1371/journal.pone.0064702.
[54] R.A. Verdugo et al., Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers, PLoS One 8 (1) (2013) e50888, http://dx.doi.org/10.1371/journal.pone.0050888.
[55] J. Jalkanen et al., Aberrant circulating levels of purinergic signaling markers are associated with several key aspects of peripheral atherosclerosis and thrombosis, Circ. Res. 116 (7) (2015) 1206–1215, http://dx.doi.org/10.1161/CIRCRESAHA.116.305715.
[56] S. Cicko et al., Purinergic receptor inhibition prevents the development of smoke-induced lung injury and emphysema, J. Immunol. 185 (1) (2010) 688–697, http://dx.doi.org/10.4049/jimmunol.0904042.
[57] B. Hechler, C. Gachet, Purinergic receptors in thrombosis and inflammation, Arterioscler. Thromb. Vasc. Biol. 35 (11) (2015) 2307–2315, http://dx.doi.org/10.1161/ATVBAHA.115.303395.
[58] H. Weidmann et al., SASH1, a new potential link between smoking and atherosclerosis, Atherosclerosis 242 (2) (2015) 571–579, http://dx.doi.org/10.1016/j.atherosclerosis.2015.08.013.
[59] Y. Zhang et al., Self-reported smoking, serum cotinine, and blood DNA methylation, Environ. Res. 146 (2016) 395–403, http://dx.doi.org/10.1016/j.envres.2016.01.026.
[60] E.S. Wan et al., Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome, Hum. Mol. Genet. 21 (13) (2012) 3073–3082, http://dx.doi.org/10.1093/hmg/dds135.
[61] F. Guida et al., Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation, Hum. Mol. Genet. 24 (8) (2015) 2349–2359, http://dx.doi.org/10.1093/hmg/ddu751.
[62] L.G. Tsaprouni et al., Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation, Epigenetics 9 (10) (2014) 1382–1396, http://dx.doi.org/10.4161/15592294.2014.969637.

Please cite this article in press as: V. Belcastro et al., The sbv IMPROVER Systems Toxicology computational challenge: Identification of human and species-independent blood response markers as predictors of smoking exposure and cessation status, Comput. Toxicol. (2017), http://dx.doi.org/10.1016/j.comtox.2017.07.004