Published in final edited form as:

Comput Toxicol. 2018 February ; 5: 31–37. doi:10.1016/j.comtox.2017.07.003.

Human blood gene signature as a marker for smoking exposure: computational approaches of the top ranked teams in the sbv IMPROVER Systems Toxicology challenge

Adi L. Tarca^{a,b,*}, Xiaofeng Gong^{c,d}, Roberto Romero^{e,f,g,h}, Wenxin Yang^{c,i}, Zhongqu Duan^{c,d}, Hao Yang^{c,i}, Chengfang Zhang^{c,d}, and Peixuan Wang^{c,d}

^aDepartment of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI, 48202, USA

^bDepartment of Computer Science, Wayne State University College of Engineering, Detroit, MI, USA

^cSJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China

^dDepartment of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

^ePerinatology Research Branch, NICHD/NIH/DHHS, Bethesda, MD, and Detroit, MI, 48201, USA

^fDepartment of Obstetrics and Gynecology, University of Michigan, Ann Arbor, MI, 48109, USA

^gDepartment of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, 48825, USA

^hCenter for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, 48201, USA

ⁱSchool of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

Abstract

Crowdsourcing has emerged as a framework to address methodological challenges in omics data analysis and assess the extent to which omics data are predictive of phenotypes of interest. The sbv IMPROVER Systems Toxicology Challenge was designed to leverage crowdsourcing to determine whether human blood gene expression levels are informative of current and past smoking. Participating teams were invited to use a training gene expression dataset to derive parsimonious models (up to 40 genes) that can accurately classify subjects into exposure groups: smokers, former smokers that quit for at least one year, and never-smokers. Teams were ranked based on two classification performance metrics evaluated on a blinded test dataset. The analytical approaches of the first- and third-ranked teams, that are presented in detail in this article, involved feature selection by moderated *t*-test or LASSO regression and linear discriminant analysis (LDA) and logistic regression classifiers, respectively. While the 12-gene signature of the top team allowed the classification of current smokers with 100% sensitivity at 93% specificity, discriminating former smokers from never-smokers was much more challenging (65% sensitivity at 57% specificity). Gene ontology molecular functions and KEGG pathways associated with

*Corresponding author: Adi L. Tarca, 3990 John R., Detroit, MI, 48201, USA; adi@wayne.edu.

current smoking included *G protein-coupled receptor activity*, *signaling receptor activity*, *calcium ion binding*, and the *Neuroactive ligand-receptor interaction* pathway. Selection of marker genes by either moderated *t*-test or multivariate LASSO regression followed by LDA or logistic regression, are robust approaches to classification with omics data, confirming in part findings of previous sbv IMPROVER challenges. While current smoking is accurately identified based on blood mRNA levels, smoking cessation for more than one year is accompanied by a “normalization” of the expression of certain mRNAs, making it difficult to distinguish former smokers from never-smokers.

Keywords

Systems toxicology; computational challenge; gene signature; smoking biomarker; predictive modeling

Introduction

Omics technologies are widespread in life science research because of their ability to measure the activity of tens of thousands of molecules (e.g., mRNAs), providing a systems-level snapshot of molecular activity in a biological sample [1]. Omics data are used to identify novel disease subtypes and markers of disease [2–5], understand the mechanisms of disease and normal physiologic processes [6,7], and more recently, to evaluate the impact of exposure to different toxicants, termed *systems toxicology* [8–10]. Given the high-dimensionality of these data and hence the ease with which data overfitting can occur [11], an important concern in the field is how to develop reliable models that relate gene expression patterns to an outcome of interest, such as exposure to a toxicant. Crowdsourcing the analysis of omics data has emerged as a framework to identify robust methodologies that address this concern [12]. The Industrial Methodology for Process Verification of Research (IMPROVER) [13] was introduced as an industry-driven initiative to determine best practices for predictive modeling with omics data, similar to the MicroArray Quality Control (MAQC)-II initiative [14] sponsored by the U.S. Food and Drug Administration. The first two sbv IMPROVER initiatives addressed the task of human disease classification from gene expression data (Diagnostic Signature Challenge) [15,16], and prediction of protein phosphorylation and pathway perturbation as responses to stimuli within and across species (Species Translation Challenge) [17–19]. The Systems Toxicology challenge, which is the subject of this article, was designed to determine whether mRNA transcription levels in blood hold information on smoking status (sub-challenge 1, SC1), and whether there is a molecular signature which is translatable from animal models to humans (sub-challenge 2, SC2). In SC1, microarray gene expression data generated from blood samples of current and former smokers, as well as from never-smokers, was provided to participating teams to develop supervised classification models [20]. The organizers provided two test datasets in which the exposure status was concealed from participants, while the identity of the teams was concealed from the organizers, who ranked the teams based on their prediction performance on the test dataset.

This article describes the approaches and results of the first- and third-ranked teams (the second-ranked team was invited to participate in the article, but declined owing to a conflict of interest) that participated in SC1, focusing on the key aspects of the methodology that led to their successful models, as well as on the similarity and differences of the resulting cigarette smoke exposure classification models.

Methods

Challenge organization

The human training dataset that was made available to participating teams was based on the Queen Anne Street Medical Centre (QASMC) clinical case-control study conducted at the Heart and Lung Centre (London, UK) according to good clinical practices (see [ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01780298), identifier NCT01780298). The dataset included gene expression data from 109 smokers, 57 former smokers, and 58 never-smokers. The test dataset was obtained by profiling blood samples from a banked repository (BioServe Biotechnologies Ltd., Beltsville, MD, USA), and included samples from 27 smokers, 26 former smokers, and 28 never-smokers. After total RNA extraction, hybridization was performed on Affymetrix GeneChip® Human Genome U133 Plus 2.0 arrays (Affymetrix, Santa Clara, CA, USA). CEL files of raw gene expression data were background corrected, normalized, and summarized into one expression value per ENTREZ Gene ID using frozen robust microarray analysis (fRMA) [21]. For the exact versions of software used, refer to the technical document from the challenge organizers [22].

Participating teams were invited to develop a prediction rule based on the training data and apply it to the test data. The test data were released in two batches to ensure that they were not used in the training processes in any way, as had happened with the second-best overall team in the sbv IMPROVER Diagnostic Signature Challenge [16]. For each test sample, the teams provided a confidence value (probability ranging from 0 to 1) that the sample was taken from a smoker (p_s) as opposed to a non-current smoker (former smoker or never-smoker) ($1-p_s$). For samples assigned to the non-current smokers group ($p_s = 0.5$), a second classification was requested to determine whether the sample came from a former smoker (p_{fs}) or from a never-smoker ($1-p_{fs}$). Of note, teams that wrongly classified non-current smokers as smokers were penalized by imputing the missing confidence values of these samples in the former smoker vs. never-smoker classification with the worst-case scenario (e.g., p_{fs} for a former smoker misclassified as current smoker was set to 0.0 instead of the ideal 1.0). An alternative evaluation of the prediction performance for this classification task of the first and third ranked teams was based on actual confidence values (requested post-challenge) to compute performance estimates that were comparable and independent between the two classification tasks (smoker vs. non-current smoker and former smoker vs. never-smoker).

The submissions from the participating teams were ranked by computing the area under the precision recall curve (AUPR) [23] and the Matthews correlation coefficient (MCC) [24] for both classification tasks on the test dataset. While the AUPR statistic monitors the precision (positive predictive value) as a function of recall (sensitivity) determined at all cut-off points on the confidence values, the calculation of the MCC statistic involves using a single

threshold on the confidence values (i.e., 0.5), hence giving a snapshot of the prediction performance at a single (fixed) sensitivity value. Each team's sum of ranks, based on the rank from each of the four statistics (two metrics $\times$ two classification tasks) was used to rank the teams. To determine whether or not the performance metric (AUPR or MCC) of a given team was better than expected by chance, the empirical distribution of the AUPR and MCC statistics was determined as described by *Belcastro* et al. in this issue. The top three teams were invited to contribute to this manuscript; the team ranked second, however, could not join owing to a conflict of interest.

Approach of the first-ranked team (Team 264)—The approach of the first-ranked team (ALT and RR, Team 264) was similar to the one used in previous sbv IMPROVER challenges [17,25], and was applied in the same manner to both classification tasks. In essence, with this approach genes are first ranked by moderated *t*-test [26] which is known to be more robust than ordinary *t*-test in selection of truly differentially expressed genes. With moderated *t*-tests gene level expression variance is shrunk toward a common value derived using information from all genes on the array. This ensures that genes are not selected as best predictors solely because they have small within group variance, and that, genes that appear to have large variance in the training data are not completely overlooked as long as they have a sizable difference in means between groups. Further, genes are filtered out regardless their significance *p*-value from the moderated *t*-test if their magnitude of change between groups does not exceed a minimum threshold, since larger differences are more likely to be reproducible in the test set. These steps are designed to remove noisy genes from the pool of candidate genes out of which an optimal number of predictors will be selected as follows:

It involved the following steps:

1. Rank genes by moderated *t*-test [26] *p*-values and select the top N_F genes (optimal value of N_F to be determined), as the genes with nominal $p < 0.05$ and fold change in expression between groups greater than a given Fold Change Threshold (*FCT*). If there are no N_F genes meeting these criteria, use (or complement with) top genes ranked solely by *p*-values. Of note, *p*-values were not adjusted for multiple testing since they were primarily used to rank the genes and not to claim significance of the association of individual genes and the phenotype.
2. Implement a three-fold cross-validation scheme repeated 10 times to estimate the cross-validated performance (P_{CV}) of a linear discriminant analysis (LDA) model using *d* genes as predictors, $d=1, 2, \dots, N_F$. The optimal value of *d* will be the one that maximizes the average performance over the 3×10 cross-validation test partitions.

In the steps above, the prediction performance was the average over five different metrics used in previous sbv IMPROVER challenges: *Belief Confusion Metric*, *Correct Class Enrichment Metric*, *Area Under the ROC curve*, *Area Under the Precision-Recall Curve*, and *Balanced Accuracy*. These metrics have been previously described [16]. The optimal values for *FCT* and N_F (leading to highest P_{CV}) were selected by trial and error over a set of pre-

specified values $FCT=(1.25, 1.5, 2.0)$ and $N_F=(5, 6, \dots, 25)$. Let p denote the optimal number of predictor genes determined using the optimal value of FCT and N_F as described in steps 1 through 2 above. A p -genes final model was fitted using the entire training dataset and then applied to the test data to compute the posterior probability that a given sample belongs to a class of interest (e.g., to the smokers group when comparing smokers against non-current smokers, or to the former smokers group when comparing former smokers against never-smokers). All analysis was performed using the *R* statistical language and environment [27] (version 3.2.2), as well as specialized packages *limma* [26] (version 3.24.15) for linear model fitting and *MASS* [28] (version 7.3–45) for LDA model fitting.

Approach of the third-ranked team (Team 259)—While the first ranked team dealt with the problem of having more candidate predictors (p) than the number of samples (n) by selecting a small subset of genes by filtering, the team SJTU (Team 259) used classification methods specifically designed for a $p \gg n$ scenario and have been used previously for microarray data analysis, including in previous IMPROVER challenges. Such approaches included Least absolute shrinkage and selection operator (LASSO) regression [29], ridge regression [30], and elastic net regression [31]. These methods are based on a logistic regression framework and differ in the level of sparsity they impose as they shrink the coefficients in the models towards 0 [31]. Since no particular model is expected to outperform all other in all situations, trial and error with different approaches is required for optimal results.

The approach of team Team 259 was applied differently for the two sub-classification tasks, as follows:

Smokers vs. non-current smokers: A multivariate logistic regression model was used as the prediction model for this classification. Least absolute shrinkage and selection operator (LASSO) regression [29], ridge regression [30], and elastic net regression [31] were used to select genes as inputs in the logistic regression model. The accuracy of the set of genes selected by each approach (having coefficients different from 0) was estimated using a split-sample procedure repeated 100 times, in which two thirds of the data were used for training and one third for testing. Since the genes identified with LASSO regression yielded the highest average accuracy when classifying smokers vs. non-current smokers, this feature selection approach was retained and used to select predictor genes for the final logistic regression model that was used on the entire training dataset and for making predictions on the test datasets.

Former smokers vs. never-smokers: For this classification task, the team proceeded similarly to the method described above; however, since the LASSO-identified genes led to poor accuracy (close to 50%), the team used a random forest algorithm [32] implemented in the *randomForest* package in *R* [27]. To reduce the number of predictors used as inputs into the random forest model, genes were ranked by nominal p -values from a t -test, and genes with $p < 0.01$ were retained. Next, a random forest model was fitted to identify genes relevant for this algorithm; those genes were used as predictors in a naïve Bayes model as implemented in the *e1071* package in *R*. Other machine-learning models were fitted and

compared using the repeated split-sample procedure described above, but none surpassed the naïve Bayes model, and hence this was the final model for this comparison.

Results

Prediction performance in the Systems Toxicology Sub-Challenge 1

Of the 23 international teams that provided submissions in SC1, only 12 met all the criteria for inclusion in the final ranking. Participating teams submitted confidence levels (ranging from 0 to 1) for classifying 27 smokers, 26 former smokers, and 28 never-smokers into smokers vs. non-current smokers (task 1), and confidence levels for classifying all non-current smokers further into former smokers vs. never-smokers (task 2). After ranking the teams by the AUPR and MCC for each classification task, the sum of ranks was computed for all teams (Table 1 and Figure 1).

The classification of subjects into smokers vs. non-current smokers was achieved with high performance, as the top three teams achieved an AUPR of 0.95–0.97, while the MCC reached 0.9 for the first ranked team (Table 1). The equivalent sensitivity of the top prediction model was 100% at 93% specificity. Additional information about the quality of the classifications not directly captured by the AUPR and MCC statistics is presented in Figure 2, which shows the distribution of confidence values for the predictions of the top three teams. Team 264 classified both actual smokers and non-current smokers as such (true positives and true negatives), with higher confidence (paired Wilcoxon test $p=1.3\times10^{-8}$) than the next-ranked two teams (Team 264 confidence median: 0.99, IQR: 0.93–1.0) (Figure 2A).

The classification of subjects into former smoker vs. never-smoker was more difficult, with the top team reaching an AUPR of 0.66 and MCC of 0.23. The equivalent sensitivity of the top prediction model was 65% at 57% specificity. The interquartile range of confidence values that a given subject was a former smoker overlapped between the actual former smokers and never-smokers for Team 264 and Team 259, but not for Team 225. However, this did not necessarily translate into a higher AUPR value for Team 225 than for Team 264, because Team 225 was penalized for classifying seven former smokers and three never-smokers as current smokers (Figure 2 B), while Team 264 only misclassified four subjects. In addition, while for Team 264, the median confidence that a sample belonged to a former smoker was above 0.5 for actual former smokers and below 0.5 for never-smokers, this was not the case for the second- and third-ranked teams, indicating model calibration issues in their models (Figure 2B).

A visualization of the training and test datasets is presented in Figure 3, using principal component analysis (PCA) based on the predictor genes identified by the top-ranked team on the training data. This figure shows that the first principal component captures variability in the expression data derived from smoking status, while the second principal component captures variability derived from the different patient populations (training vs. test sets).

Analysis of the predictor genes identified by the first- and third-ranked teams

Table 2 shows the symbols of predictor genes and their weight in the two LDA models of Team 264, designed to discriminate between smokers and non-current smokers (task 1), as

well as between former smokers and never-smokers (task 2). For each of these two classifications, the fRMA summarized expression level (on a log 2 scale) can be multiplied with the corresponding coefficients and summed. If the weighted gene expression score is greater than the determined threshold, then the subject is assigned to the positive group in each comparison (i.e., smoker for task 1 and former smoker for task 2). Similarly, Table 3 shows the predictor genes used by Team 259 in a logistic regression and naïve Bayes classifier for tasks 1 and 2, respectively.

An over-representation analysis conducted on the union of predictor genes of the first- and third-ranked teams (Tables 2 and 3) identified Gene Ontology Consortium [33] molecular functions *G protein-coupled receptor activity*, *signaling receptor activity*, *calcium ion binding*, and the Kyoto Encyclopedia of Genes and Genomes (KEGG) *Neuroactive ligand-receptor interaction* pathway enriched among the predictor genes that discriminated smokers from non-current smokers (q-values<0.05, odds ratios ranging from 4.2 to 21.0).

Discussion

Predictive analytics with omics data is challenging and plagued with pitfalls, as there is typically a low ratio of independent observations (samples, n) to variables measured (genes, p), also known as an $n \ll p$ problem. Standard statistical pattern recognition models (e.g., logistic regression, LDA) cannot be fitted in this setup, while machine-learning algorithms that could be applied (e.g., decision trees, neural networks) would not perform optimally as most of the variables are not informative of the outcome, and reliable parameters in these models cannot be properly estimated. This in turn can lead to classification models that accurately classify the training dataset but do not generalize when applied to new datasets. Therefore, there is a need in the field to find analysis approaches that are both practical and robust, and demonstrate that the observed associations between gene expression data and a given phenotype are reproducible. Crowdsourcing emerged as a solution to these computational needs, by exploring the current knowledgebase and unbiasedly comparing existing analytic alternatives, while at the same time demonstrating that a real signal in the data exists. This latter goal is demonstrated when models generated by competing teams on training datasets predict better than expected by chance on an independent analyst-blinded dataset. Below, we discuss to what extent these two goals were met in SC1 of the sbv IMPROVER Systems Toxicology challenge.

From a computational perspective, SC1 confirmed the robustness of the classification pipeline proposed by Tarca et al. [16,17,25], which was ranked at the top in the previous two sbv IMPROVER challenges, and was ranked first in the current sub-challenge. The prediction performance of Team 264 was significantly better than the one of the other teams, and was based on selection of few predictor genes used as inputs in an LDA model. However, it is not feasible to pinpoint which were the modeling factors that made the difference in the team performance data reported in Table 1. In the smoker vs. non-current smoker classification task, the first- and third-ranked teams achieved similar AUPR values, but the MCC of the first-ranked team was higher by 0.11. The first-ranked team used 12 genes as predictors, while the third-ranked team used 14 genes; eight genes were common to both groups. Different methods were used to select these genes (moderated t -test gene

ranking vs. multivariate LASSO regression), and different performance metrics were considered to guide the search for the optimal set of genes (average over five different metrics, vs. accuracy). In addition, different procedures were used to estimate the performance metrics (repeated cross-validation vs. repeated split-sample). Of note, the exact prediction performance metric used for team ranking on the test dataset was not disclosed to the participating teams in the Systems Toxicology challenge, and hence each team sought to generate a model that optimized the performance metric of their choice. As we have previously shown, different modeling pipelines may be required to maximize a particular prediction performance metric [16].

The main biological finding of SC1 of the sbv IMPROVER Systems Toxicology challenge is that smoking exerts a significant effect on the blood transcriptome, and that this effect weakens after one or more years of smoking cessation. A possible explanation for the differential regulation of gene expression due to smoking is the fact that cigarette smoking has a broad and long-term impact on genome-wide methylation [34], while this epigenetic mechanism is known to allow the organism to respond to the environment through changes in gene expression [35]. Although the prediction models to classify smokers vs. non-current smokers obtained by the first- and third-ranked teams were based on 12 and 14 predictor genes, respectively, these are not necessarily all the differentially expressed genes due to smoking, and neither are all of these genes necessarily differentially expressed using typical differential expression criteria. A differential expression analysis based on moderated *t*-tests [26] and false discovery rate (FDR) controlled at 10% would yield 51 genes differentially expressed in current smokers, based on the training set that would meet the same significance criteria on the test set while being concordant in terms of direction of change.

Although much reduced, the accuracy of prediction of past smoking was still better than expected by chance. For Team 264 all seven predictor genes selected to discriminate between former smokers and never-smokers (Table 2) changed in the same direction in the test dataset as they did in the training dataset. However, a typical differential expression analysis would fail to find any gene significant at 10% FDR in both the training and test datasets, highlighting why sometimes even controlling the FDR at 10% can be conservative for microarray studies, as we have shown elsewhere [36].

In conclusion, sub-challenge 2 of the IMPROVER Systems Toxicology challenge has demonstrated that gene expression-based classification approaches identified previously as top contenders (moderated-test selection of few genes + LDA classifier)[16,17,25] proved yet again to be robust and achieve similar performance with fewer predictor genes than more complex approaches designed for high-dimensional data (LASSO regression + random forests). Future crowdsourcing initiatives assessing translation of findings between species should enforce consistency in the direction of change of genes across species in order to maximize the chance of finding the best *species-independent* gene signatures.

Abbreviations

KEGG Kyoto Encyclopedia of Genes and Genomes

IMPROVER Industrial Methodology for PROcess VERification in Research

LDA	Linear Discriminant Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
sbv	systems biology verification

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995; 270:467–470. [PubMed: 7569999]
2. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*. 1999; 96:6745–6750. [PubMed: 10359783]
3. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet*. 2000; 24:227–235. [PubMed: 10700174]
4. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000; 403:503–511. [PubMed: 10676951]
5. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol*. 2006; 195:373–388. [PubMed: 16890548]
6. Hassan SS, Romero R, Tarca AL, Draghici S, Pineles B, et al. Signature pathways identified from gene expression profiles in the human uterine cervix before and after spontaneous term parturition. *Am J Obstet Gynecol*. 2007; 197:250 e251–257. [PubMed: 17826407]
7. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003; 34:267–273.
8. Ahuja V, Sharma S. Drug safety testing paradigm, current progress and future challenges: an overview. *J Appl Toxicol*. 2014; 34:576–594. [PubMed: 24777877]
9. Chen S, Xuan J, Couch L, Iyer A, Wu Y, et al. Sertraline induces endoplasmic reticulum stress in hepatic cells. *Toxicology*. 2014; 322C:78–88.
10. Sturla SJ, Boobis AR, FitzGerald RE, Hoeng J, Kavlock RJ, et al. Systems toxicology: from basic research to risk assessment. *Chem Res Toxicol*. 2014; 27:314–329. [PubMed: 24446777]
11. Ioannidis JP. Biomarker failures. *Clin Chem*. 2013; 59:202–204. [PubMed: 22997282]
12. Saez-Rodriguez J, Costello JC, Friend SH, Kellen MR, Mangravite L, et al. Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nat Rev Genet*. 2016; 17:470–486. [PubMed: 27418159]
13. Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, et al. Verification of systems biology research in the age of collaborative competition. *Nat Biotechnol*. 2011; 29:811–815. [PubMed: 21904331]
14. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat Biotechnol*. 2010; 28:827–838.
15. Meyer P, Hoeng J, Rice JJ, Norel R, Sprengel J, et al. Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. *Bioinformatics*. 2012; 28:1193–1201. [PubMed: 22423044]
16. Tarca AL, Lauria M, Unger M, Bilal E, Boue S, et al. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics*. 2013; 29:2892–2899. [PubMed: 23966112]
17. Dayarian A, Romero R, Wang Z, Biehl M, Bilal E, et al. Predicting protein phosphorylation from gene expression: top methods from the IMPROVER Species Translation Challenge. *Bioinformatics*. 2014
18. Rhrissorrakrai K, Belcastro V, Bilal E, Norel R, Poussin C, et al. Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge. *Bioinformatics* under review. 2014

19. Hafemeister C, Romero R, Bilal E, Meyer P, Norel R, et al. Inter-species pathway perturbation prediction via data-driven detection of functional homology. *Bioinformatics*. 2015; 31:501–508. [PubMed: 25150249]
20. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. *PLoS Comput Biol*. 2007; 3:e116. [PubMed: 17604446]
21. McCall MN, Bolstad BM, Irizarry RA. Frozen robust multiarray analysis (fRMA). *Biostatistics*. 2010; 11:242–253. [PubMed: 20097884]
22. SBV IMPROVER P. The Systems Toxicology Challenge. 2015
23. Schein, AI., Popescul, A., Ungar, LH., Pennock, DM. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Tampere, Finland: ACM; 2002. Methods and metrics for cold-start recommendations; p. 253-260.
24. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BiochimBiophysActa*. 1975; 405:442–451.
25. Tarca AL, Than NG, Romero R. Methodological approach from the Best Overall Team in the sbv IMPROVER Diagnostic Signature Challenge. *Systems Biomedicine*. 2013; 1
26. Smyth, GK. Limma: linear models for microarray data. In: Gentleman, R.Carey, VJ.Huber, W.Irizarry, RA., Dudoit, S., editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer; 2012. p. 397-420.
27. Team RDC. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2009.
28. Venables, WN., Ripley, BD. *Modern Applied Statistics with S*. New York: Springer; 2002.
29. Tibshirani R. Regression shrinkage and selection via the lasso. *JRoyStatistSocSerB*. 1996; 58:267–288.
30. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *JStatSoftw*. 2010; 33:1–22.
31. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67:301–320.
32. Breiman L. Random Forests. *Machine Learning*. 2001; 45:5–32.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *NatGenet*. 2000; 25:25–29.
34. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet*. 2016
35. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003; 33(Suppl):245–254. [PubMed: 12610534]
36. Pappas A, Chaiworapongsa T, Romero R, Korzeniewski SJ, Cortez JC, et al. Transcriptomics of Maternal and Fetal Membranes Can Discriminate between Gestational-Age Matched Preterm Neonates with and without Cognitive Impairment Diagnosed at 18–24 Months. *PLoS One*. 2015; 10:e0118573. [PubMed: 25822971]

Highlights

- Parsimonious blood gene expression signatures accurately predict current smoking.
- Effect of smoking on blood mRNA expression is reduced after 1 year of cessation.
- Best predictive methods found in previous challenges are proven again to be robust.

Figure 1. Classification performance of 12 teams with valid submissions in the sbv IMPROVER Systems Toxicology sub-challenge 1

Data shown represent the ranks (1–12, the smaller the better) for two prediction performance metrics (area under the precision-recall curve, AUPR, and Matthews correlation coefficient, MCC) in two classification tasks (smoker vs. non-current smoker and former smoker vs. never-smoker). The final team ranking is based on the sum of the four individual ranks.

Figure 2. Classification confidence values for the top three teams in sub-challenge 1

Data shown represent the confidence (0.0–1.0) that blood gene expression profiles belonged to a smoker (A, left) or former smoker (B, right). Distribution boxplots are shown by actual smoking status. For team 225 in panel B, data for 10 subjects are missing as this team misclassified seven former smokers and three never smokers as current smokers.

Figure 3. Principal component analysis plot of the training and test sets based on the predictor genes identified by the top-ranked team (Team 264) on the training set

The figure shows expression profiles reduced to the first two principal components for the training (left) and test (right) datasets involved in the smoker vs. non-current smoker (top) and former smoker vs. never smoker (bottom) classifications. Samples correctly classified by Team 264 are shown as bullets, while misclassified samples are shown as triangles. Sens., sensitivity; Spec., specificity; PC, principal component.

Table 1
Performance metrics for participating teams in sub-challenge 1 of the sbv IMPROVER Systems Toxicology challenge

Teams were ranked by the area under the precision recall curve (AUPR) and the Matthews correlation coefficient (MCC) in each of the two classification tasks. The ranks of teams (the smaller the better) on each individual metric and task were summed to determine the final rank. The AUPR and MCC values displayed in parentheses for the smoker vs. never-smoker classification were determined from all actual confidence values, as opposed to the values outside the parentheses (used in the official ranking), in which missing predictions for some non-current smokers (wrongly classified as smokers) were imputed with the worst-case scenario value (see the Methods section). This explains why even apparently worse than chance performance values (e.g., AUPR<0.5 or MCC<0.0 for Team 259) appear as significantly better than chance ($p<0.05$) based on simulating the null distribution of these statistics.

Team	Final rank	Sum ranks	Smoker vs. Non-current Smoker				Former Smoker vs. Never-Smoker			
			AUPR	MCC	p AUPR	p MCC	AUPR	MCC	p AUPR	p MCC
264	1	5	0.96	0.90	0.000	0.000	0.58 (0.66)	0.07 (0.23)	0.000	0.000
225	2	10	0.97	0.77	0.000	0.000	0.50 (NA)	0.02 (NA)	0.000	0.000
259	3	15	0.95	0.79	0.000	0.000	0.47 (0.58)	-0.02 (0.12)	0.000	0.000
221	4	20	0.79	0.78	0.000	0.000	0.48	-0.01	0.000	0.000
247	5	20	0.90	0.67	0.000	0.000	0.45	0.04	0.000	0.000
269	6	25	0.91	0.71	0.000	0.000	0.42	-0.11	0.000	0.000
283	7	28	0.86	0.63	0.000	0.000	0.48	-0.11	0.000	0.001
250	8	29	0.85	0.67	0.000	0.000	0.43	-0.07	0.000	0.000
222	9	34.5	0.87	0.67	0.000	0.000	0.35	-0.32	0.005	0.062
290	10	35.5	0.83	0.59	0.000	0.000	0.46	-0.30	0.000	0.057
257	11	45	0.23	-0.05	1.000	0.705	0.29	-0.80	1.000	0.997
215	12	45	0.21	-0.50	1.000	1.000	0.30	-0.79	0.982	0.997

Table 2
Gene predictors and corresponding weights in the linear discriminant analysis models of Team 264

For each classification task, weighted gene expression levels by corresponding coefficients are summed; if the result is higher than the threshold value, the subject is classified as a smoker (task 1, left panel) or as a former smoker (task 2, right panel).

Smoker vs. Non-current Smoker		Former Smoker vs. Never-Smoker	
Gene	Coefficient	Gene	Coefficient
LRRN3	0.6655	SLC38A3	1.8199
AHRR	0.2526	TMEM246	1.0501
LINC00599	-0.0763	HSF4	0.6945
SASH1	0.5862	SEPT2	-0.7158
P2RY6	0.8754	PBK	-0.8741
CTTNBP2	0.1618	PRR29	0.9786
GPR15	1.2430	POU4F1	-3.1714
SEMA6B	0.4604	<i>Threshold</i>	2.5978
CDKN1C	-1.4433		
PID1	0.1891		
CLEC10A	0.9486		
F2R	-0.3384		
<i>Threshold</i>	19.7082		

Table 3
Gene predictors used in the models of Team 259

For the first classification task (smoker vs. non-current smoker, left panel), gene expression levels are multiplied to corresponding coefficients of a logistic regression model and added to the intercept term. If the result is >0 , the subject is classified as smoker; otherwise, the subject is classified as a non-current smoker. For the second classification task, a naïve Bayes classifier was used instead of logistic regression, and hence a simple classification formula is not available.

Smoker vs. Non-current Smoker		Former Smoker vs. Never-Smoker	
Gene	Coefficient	Gene	
CDKN1C	-1.194	CT83	SLC17A5
DSC2	0.306	POU4F1	CCDC138
GPR15	0.986	LOC648691	POU5F1P4
P2RY6	0.783	ZNF280B	KSR1
PTGFR	0.162	KCNE3	PTPN3
CLEC10A	0.684	NLRP8	
FSTL1	-0.258	SLC17A7	
GSE1	-0.486	<i>NS3BP</i>	
SASH1	0.577	HSF4	
LRRN3	1.012	KIAA1522	
PID1	0.044	FAM170A	
AHRR	0.749	ZNF430	
GPR63	0.795	LOC90246	
CRACR2B	0.210	PTPN14	
<i>Intercept</i>	-22.906	OLIG2	