BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 22 2013, pages 2892–2899
doi:10.1093/bioinformatics/btt492

Systems biology

Advance Access publication August 20, 2013

Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge

Adi L. Tarca[1,2], Mario Lauria[3], Michael Unger[4], Erhan Bilal[5], Stephanie Boue[6], Kushal Kumar Dey[4], Julia Hoeng[6], Heinz Koeppl[4], Florian Martin[6], Pablo Meyer[5], Preetam Nandy[4], Raquel Norel[5], Manuel Peitsch[6], Jeremy J. Rice[5], Roberto Romero[2], Gustavo Stolovitzky[5,*], Marja Talikka[6], Yang Xiang[6], Christoph Zechner[4] and IMPROVER DSC Collaborators

[1]Department of Computer Science, Wayne State University, [2]Perinatology Research Branch, NICHD/NIH, Detroit, MI 48201, USA, [3]The Microsoft Research - University of Trento Centre for Computational and Systems Biology, Rovereto 38068, Italy, [4]ETH Zurich, Zurich 8092, Switzerland, [5]IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA and [6]Philip Morris International, Research & Development, Neuchâtel CH-2000, Switzerland

Associate Editor: Martin Bishop

ABSTRACT

Motivation: After more than a decade since microarrays were used to predict phenotype of biological samples, real-life applications for disease screening and identification of patients who would best benefit from treatment are still emerging. The interest of the scientific community in identifying best approaches to develop such prediction models was reaffirmed in a competition style international collaboration called IMPROVER Diagnostic Signature Challenge whose results we describe herein.

Results: Fifty-four teams used public data to develop prediction models in four disease areas including multiple sclerosis, lung cancer, psoriasis and chronic obstructive pulmonary disease, and made predictions on blinded new data that we generated. Teams were scored using three metrics that captured various aspects of the quality of predictions, and best performers were awarded. This article presents the challenge results and introduces to the community the approaches of the best overall three performers, as well as an R package that implements the approach of the best overall team.

The analyses of model performance data submitted in the challenge as well as additional simulations that we have performed revealed that (i) the quality of predictions depends more on the disease endpoint than on the particular approaches used in the challenge; (ii) the most important modeling factor (e.g. data preprocessing, feature selection and classifier type) is problem dependent; and (iii) for optimal results datasets and methods have to be carefully matched. Biomedical factors such as the disease severity and confidence in diagnostic were found to be associated with the misclassification rates across the different teams.

Availability: The lung cancer dataset is available from Gene Expression Omnibus (accession, GSE43580). The *maPredictDSC* R package implementing the approach of the best overall team is available at www.bioconductor.org or http://bioinformaticsprb.med.wayne.edu/.

Contact: gustavo@us.ibm.com

*To whom correspondence should be addressed.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 8, 2013; revised on August 4, 2013; accepted on August 16, 2013

1 INTRODUCTION

Microarrays were introduced in life science research as a practical means to measure whole-genome expression levels (Schena *et al.*, 1995). Typical experiments involving microarray technologies were designed to gain biological insights into various conditions but also to discover new, and predict predefined, disease phenotypes. For instance, breast tumors were classified based on their molecular profiles more than a decade ago (Bittner *et al.*, 2000; Gordon *et al.*, 2002; Perou *et al.*, 2000), and progress has been steady and promising; yet, practical applications of microarrays in patient care are only emerging. A recent comparison of three microarray-based classifiers for breast cancer subtyping, BluePrint, MammaPrint and TargetPrint, concluded that multigene assays were more reliable than clinicopathological criteria alone for the clinical management of breast cancer patients (Nguyen *et al.*, 2012). Pharmacogenetic (PGx) testing is an essential part of personalized medicine, which aims to predict an individual's risk for adverse drug response or treatment outcome. Recently, Hresko and Haga (2012) reviewed how PGx tests are accepted in the US in terms of coverage by the largest health insurance companies. Although there is no US Food and Drug Administration (FDA)-approved test available for OncotypeDx [a 21-gene assay that can predict 10-year distant breast cancer recurrence (Kaklamani, 2006; Yamani *et al.*, 2007a)], it is covered by most major health insurances. AlloMap [a 20-gene test that can assess the risk of cardiac allograft rejection following a heart transplant (Mook *et al.*, 2007; Yamani *et al.*, 2007b)], similarly lacks FDA approval, but it is considered beneficial by a number of health insurances. In contrast, although the MammaPrint test

[70-gene profile for prognostic and predictive tumor analysis (Acharya *et al*., 2012)] is approved by FDA, it has rather limited insurance coverage. Apparently, FDA approval is neither a prerequisite nor sufficient for the acceptance of a given test; rather, independent review by insurance companies and confidence by the clinicians (and patients) may be the main factors that influence the uptake of a new PGx test. We postulate that better and more standard methods to verify diagnostic PGx tests would facilitate acceptance by regulatory agencies and healthcare providers and hasten deployment to the public.

Recently, the Industrial Methodology for PROcess VErification in Research (IMPROVER) was designed as a methodology to validate industrial research processes related to systems biology (Meyer *et al*., 2011). As a first initiative of the IMPROVER project, the Diagnostic Signature Challenge (DSC) (Meyer *et al*., 2012) was designed to determine to what extent transcriptomic data can be used for phenotype prediction and to test which computational approach works best for this end. Participants in the challenge were asked to produce a prediction model (classifier) that can infer the phenotype of biological samples from gene expression data for five different endpoints. The teams were ranked based on the prediction performance on test datasets generated by the organizers (for details see the Methods section).

The purpose of this article is threefold: one, to describe the IMPROVER DSC results including classification performance on each endpoint, scoring methodology and overall ranking of the teams as well as ranking stability with respect to the composition of the test datasets; two, to introduce the methods of the top three overall performers and discuss the performance of an ensemble classifier that aggregates the predictions from best models submitted in the challenge; and three, to identify some of the sources of variability in the classification performance, including biomedical factors, and point toward the best alternatives at each step in the classification pipeline. Performance data from the models submitted to the IMPROVER DSC and from post-challenge computational studies are used to support our findings. We conclude this article with a summary of the main observations and discuss their relevance.

2 METHODS

2.1 Organization of the challenge

A full description of the DSC is available at www.sbvimprover.com. Briefly, participants in the challenge were asked to produce a prediction model that can infer the phenotype of biological samples from microarray data. Although no constraints were placed on the type of approach to be used, all participants used different flavors of machine learning algorithms (Tarca *et al*., 2007) trained on public microarray datasets. The training datasets for each sub-challenge can be downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/gds/?term=GEO) or from ArrayExpress at European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/arrayexpress/). These datasets represent gene expression levels profiled using commercial microarrays platforms including Affymetrix (Santa Clara, CA) and Illumina (San Diego, CA) platforms. Details on each sub-challenge including the nature and composition of the test datasets and the accession numbers of the suggested training datasets are given as Supplementary Information. The phenotypes/endpoints considered in the challenge included psoriasis, multiple sclerosis (MS) Diagnostic and MS Stage, lung cancer (LC) and chronic obstructive pulmonary disease (COPD). Each phenotype/endpoint was treated as a stand-alone sub-challenge. For a number of reasons, the difficulty in predicting each of these endpoints varied. For instance, the outcome should be easier to predict when the microarray profile was measured in the primary diseased tissue (e.g. skin in psoriasis and tumor cells in LC) as opposed to in a surrogate tissue (e.g. blood in MS). The challenge organizers provided a test dataset for each of the sub-challenges that could be used only within the context of the challenge. Each participating team in the DSC had to submit (i) the results generated as part of the DSC consisting of one belief value per sample and class in the range between 0 and 1, and (ii) a description of the algorithm(s), methodology and/or software used in sufficient detail to allow external parties to understand the basic functionality of the method.

2.2 Scoring procedures and metrics

Submissions were scored on the basis of their predictive ability with oversight by an independent scoring review panel of qualified experts (panel members are listed at http://www.sbvimprover.com/scoring). The scoring team was blind to the identity of the participating teams. The predictions from each team were required to be in the form of a *belief* value per test sample and group/class in each of the five sub-challenges. The belief values, summing up to 1.0 for each sample, represent the degree of confidence of the modeler that the test sample belonged to a given phenotype. To score the predictions from each team, the organizers chose a set of scoring metrics, some of which were specially designed for this challenge. The metrics were chosen to be non-redundant, applicable to multiclass problems, include both threshold-based and threshold-free instantiations and make use of continuous confidence levels for each class. The three metrics chosen were *Area Under the Precision-Recall Curve* (AUPR), *Belief Confusion Metric* (BCM) and *Correct Class Enrichment Metric* (CCEM). For more details of team scoring and ranking see Supplementary Information. The overall ranking of teams was based on the sum of the ranks over the endpoints for which at least one team did better than expected by chance, as determined by transforming the *Z*-scores into *P*-values and adjusting them using the false discovery rate method (Benjamini and Hochberg, 1995).

2.3 Resources availability

All test datasets used in the IMPROVER DSC were made available to the challenge participants from the project Web site (www.sbvimprover.com). The gold standard files giving the correct class labels of samples were also made public after the challenge was completed. The LC dataset is available from the National Center for Biotechnology Information GEO (GEO accession number: GSE43580).

A Bioconductor (www.bioconductor.org) compliant R package called *maPredictDSC* was created to implement the approach of the best overall team with extra capabilities for testing additional 26 combinations of methods described in the Results section, as well as aggregating the predictions from some or all the combinations explored. The three metrics used in the DSC to score the teams were also implemented in this package, which is available from http://bioinformaticsprb.med.wayne.edu and www.bioconductor.org.

3 RESULTS

3.1 Participation, scoring and team ranking in the IMPROVER DSC

The IMPROVER DSC was open from March 5 to June 21, 2012. During this time, 54 teams from around the world (Supplementary Fig. S1) participated in the competition; 34 of them made a submission for all five endpoints. The fact that many teams submitted to all sub-challenges allowed us to explore

whether the same methodology performed similarly for different endpoints. For a more detailed description of the challenge organization and endpoints characterization see the Methods section. Figure 1a shows the distribution of the three performance indices used in the challenge as a function of the endpoints. For comparison purposes, the Matthews correlation coefficient (MCC) (Matthews, 1975) used previously (Baldi *et al.*, 2000) to assess the accuracy of prediction algorithms for classification is also included in this figure. The MCC was not used in the team ranking as it cannot be applied to multiclass problems such as the LC sub-challenge. Across the four binary endpoints and all teams, the correlations between MCC and CCEM, BCM and AUPR were 0.99, 0.97 and 0.96, respectively, suggesting that MCC and CCEM are similar. Clearly, there were large differences in the quality of predictions between the endpoints, with psoriasis being the easiest to predict, followed by LC, COPD and MS Diagnostic (Fig. 1a). For the MS Stage sub-challenge, no team's performance was better than chance (after multiple testing correction); therefore, to reduce the impact of random chance on a team's ranking, this sub-challenge was not used in the final rankings. In the four-way LC classification, only the tumor type segregation, adenocarcinoma (AC) versus squamous cell carcinoma (SCC), was successfully achieved but not between different cancer stages (AC1 versus AC2 or SCC1 versus SCC2, postfixes 1 and 2 indicate the cancer stages) (see Supplementary Fig. S2). The data displayed in Figure 1a were first transformed into *Z*-scores (see Supplementary Information) so that the performances of the classifiers on a two-class endpoint (e.g. psoriasis) and a four-class endpoint (i.e. LC) could be compared. The likely biological interpretation of these results can be that the predictions in the easiest sub-challenge, psoriasis, were based on transcription levels measured in the primary diseased tissue (skin), whereas predictions in the most difficult two sub-challenges, MS Stage and MS Diagnostic, were based on transcription levels measured in a surrogate tissue, namely, blood.

A schematic representation of the overall team ranking with the best overall three teams highlighted is shown in Figure 1b. Supplementary Table S1 gives the identity and affiliations of the members of the teams that attained best performances overall and best in each of the sub-challenges of the IMPROVER DSC. The overall team ranking is provided in Supplementary Table S2, whereas ranking performance data on each individual sub-challenge is available at http://www.sbvimprover.com/. The sum of ranks over the different metrics and endpoints was chosen to rank the teams in the IMPROVER competition because it is robust to outlier values for a particular metric and/or particular endpoint, yet this measure is insensitive to *how much* better a team scored compared with another one (as long as it scored better). On the other hand, the aggregated (summed) *Z*-scores over the different metrics and endpoints (see Fig. 1c) is monotonically related to the performance on any of the metrics and endpoints, but it can be skewed by outlier values. The best and second best overall teams were the same using both the sum of ranks and aggregated *Z*-scores.

To determine the stability of the team rankings with respect to the composition of the test datasets, we performed a simulation by taking one bootstrap sample from each test dataset and computed the performance metrics and ranking again for all teams.

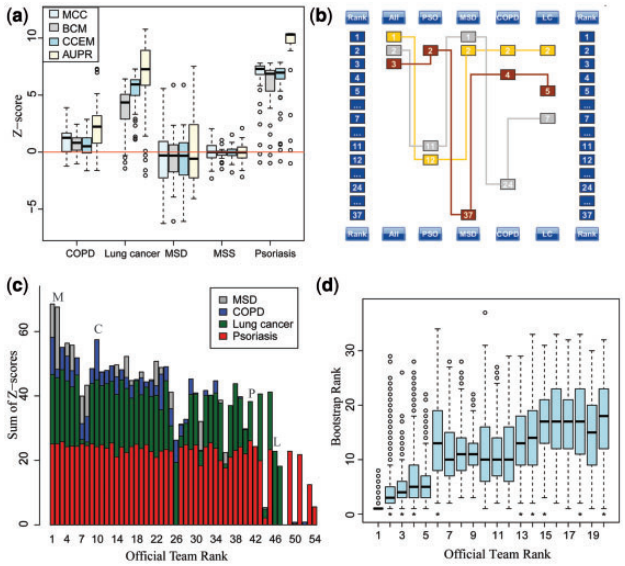

Fig. 1. (**a**) Distribution of team performance (after *Z*-score conversion) for all endpoints and metrics. (**b**) Ranks of the top three best overall teams in each sub-challenge; (**c**) sum of BCM, CCEM and AUPR (after *Z*-scores conversion) for each team. Teams are sorted in the official team ranking order. Label C marks best team for COPD, P for psoriasis, L for LC and M for MSD (MS Diagnostic). (**d**) Distribution of bootstrap ranks for each team sorted by official rank. Only data from the top 20 teams are shown. *Median values are significantly worse (one-tailed paired Wilcoxon test $P<0.05$) compared with the team ranked one position higher. MSD, MS Diagnostic, PSO, psoriasis

The distribution of the overall team ranks and the corresponding aggregated *Z*-scores were determined over 1000 such simulations and are shown in Figure 1d and Supplementary Figure S3, respectively. The distribution of the bootstrap rank values suggested that the team ranking was stable up to the fourth ranked team, because the second, third and fourth ranked teams each had a median rank worse than the team officially ranked one position better (one-tailed paired Wilcoxon test $P <$ 0.05). A stability analysis of the team ranking based on the aggregated *Z*-scores showed also that the ranking of the first and second best overall teams was stable (Supplementary Fig. S3). The performance of the top three best overall teams was again evaluated in a post-challenge swap analysis in which the training and test datasets were interchanged. The performance values in the swap analysis were, on average, only slightly worse than in the competition (see Supplementary Fig. S4).

3.2 Description of the classification pipelines of the three best overall teams

One of the main goals of the IMPROVER DSC was to find the best classification pipeline for outcome prediction based on microarray data. The approaches developed by the three best overall teams (first team 221; second team 227; and third team 161) are summarized below:

3.2.1 Team 221 Preprocess training and test datasets together using Robust Multi-array Average (RMA) (Irizarry *et al.*, 2003). From the pool of all genes expressed above the background, select a small number of differentially expressed genes using a

moderated *t*-test, *P*-value and fold change. Optimize the number of features to be included in the model by maximizing the cross-validated performance (Area Under the Receiver Operating Characteristic Curve statistic) of a linear discriminant analysis (LDA). Fit the LDA model and make predictions assuming equal priors of each class in the test set. Optimize the choice of training datasets for some endpoints, as well as gene ranking strategies and *P*-value/fold change cut-offs by trial and error. For more details see reference (Tarca *et al.*, 2013).

3.2.2 Team 227 Preprocess training and test datasets either together or separately using RMA or the Affymetrix Microarray Suite 5.0 (MAS5) algorithm. Preselect genes using a Wilcoxon test between classes on a training dataset. Build a rank-based signature for each sample in the test dataset by selecting the top- and bottom-ranked genes among the preselected genes. Determine distances between each pair of signatures using a gene set enrichment analysis -based metric. Apply unsupervised clustering of test samples using Cytoscape (Shannon *et al.*, 2003). Optimize signature size by maximizing cluster separation. Assign a class label to each cluster using the direction of expression changes for selected disease genes derived from the literature. For more details see reference (Lauria, 2013).

3.2.3 Team 161 Preprocess training and test datasets together using RMA. Select genes using a Wilcoxon test. Use the Least Absolute Shrinkage and Selection Operator (LASSO) regularized logistic regression model as the classifier (Tibshirani, 1996). Optimize the regularization parameter value by maximizing the cross-validated performance of the model on the training dataset. Choose training datasets for some of the endpoints by trial and error. For more details see reference (Nandy *et al.*, 2013).

3.3 Combining predictions from multiple classification pipelines

It has been shown previously (Marbach *et al.*, 2012; Prill *et al.*, 2010) that by combining predictions from different classifiers, the quality and robustness of predictions can be improved. The models of the various teams were different because they were often trained on different datasets using different preprocessing methods, feature selection methods and classifier types. To combine the predictions from multiple models, the samples from each endpoint test dataset were split randomly into two equal and balanced parts. The first half was used to rank the models/teams. The second half was used to assess the quality of an ensemble classifier that combined the predictions from the best n ($n = 2, \ldots, 34$) models/teams. The Z-scores for a given metric (e.g. BCM) were used to rank the teams, and then the belief matrix of the ensemble classifier was constructed by averaging the individual belief matrices from the best n models. The entire process was repeated 500 times and average performance values that were obtained are shown in Figure 2. The BCM values tended to be negatively impacted by prediction aggregation, especially when the subsequent models were much worse than the best model, which was the case for MS Diagnostic (Fig. 2). In contrast, the AUPR values seemed to benefit substantially from prediction aggregation regardless of whether the subsequent models were about the same or worse than the best model. Of note, the AUPR of 1 ($Z = 6.5$) achieved for psoriasis did not change as more models were added into the ensemble. The reason for this is that a perfect precision *versus* recall curve was already achieved by the top 12 models, and the next 13 models had an AUPR > 0.99. These findings highlight the differences between these two metrics and also show that the best predictions (single best or ensemble classifier) are metric- and endpoint dependent. In all cases (with the exception of MS Diagnostic measured by the BCM and CCEM metric, see Supplementary Fig. S5 for CCEM), the aggregate prediction performance was robust to the inclusion of poorly performing methods. This finding is especially important, as the outcome of the classification is not known in advance; therefore, the aggregation of predictions from several methods is a safe and robust strategy, because which of the individual methods performs best is not known.

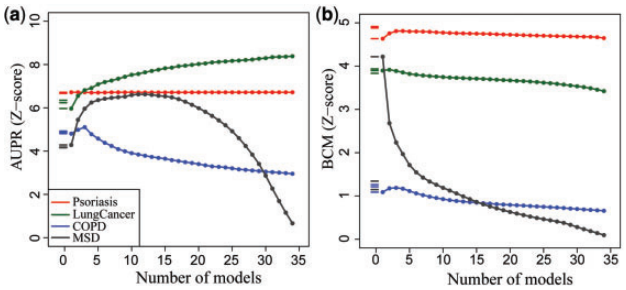

Fig. 2. Aggregation of the predictions from the best models. (**a**) AUPR and (**b**) BCM Z-scores (average over 500 trials) as a function of the number of best models used in the ensemble classifier. In each trial, half of the test data were used to rank the models and the remaining half were used to estimate the performance of the ensemble model. The averaged Z-scores for the top three individual models are shown as horizontal line segments on the left side of the graphs. MSD, MS Diagnostic

3.4 Explaining the variability in classifier performance

The ideal method of determining the contribution of each modeling factor to the overall performance of the classification pipeline would be a factorial experiment in which performance data are determined for every combination of factors (e.g. classifier type or feature selection method) in the model. The performance data obtained from the models submitted in the DSC were not amenable for this type of analysis because the types of methods that were used varied widely between teams and even the training datasets varied from one team to another. In addition, the descriptions of the methods used in the models (at the time of submission each team was required to submit a write-up describing their method) were often not sufficient to determine which method was used in each of the steps of the classification pipeline. Together these realities made it impossible to carry out a multivariate analysis of variance (ANOVA) of the performance data. Therefore, we applied a series of sub-analyses on the performance data as follows:

A two-way ANOVA analysis was applied on the model performance data (after Z-score transformation) using data from the models of the 34 teams that made a complete submission. Two factors considered were the team and the endpoint. The endpoint explained 60, 70 and 77% of the variance of AUPR, BCM and CCEM, respectively ($P < 0.05$), whereas the

team factor only explained 9, 8 and 6% of the variance, respectively (non-significant). The between-endpoint differences in performances were, therefore, much larger than the between-team differences for the same endpoint as revealed in Figure 1a.

To further isolate the contribution of each modeling factor to the variability of a model's performance, we conducted a factorial experiment based on the classification pipeline and training datasets used by the best overall team. In the factorial analysis, we considered three preprocessing methods (MAS5, RMA and GCRMA), three feature selection methods (*t*-test, moderated *t*-test and Wilcoxon test) and three different classifiers (LDA, k-Nearest Neighbor (kNN) and Support Vector Machines (SVM)). These modeling factor levels were selected from the options that were used most often in the challenge. We applied the 27 different possible combinations of the modeling factor levels to the four endpoints in the IMPROVER challenge, after converting the four-class LC sub-challenge into a two-class sub-challenge (i.e. AC versus SCC regardless of the stage) to make it possible to use identical feature selection methods across all the datasets. Performance results averaged over all four endpoints, and the three metrics are shown in Figure 3. An ANOVA of these data revealed that the endpoint was the dominant source of variability, regardless of the performance metric that was used (Supplementary Fig. S6). As an average over all three performance metrics, the endpoint explained 65% of the variability ($P<0.05$), the pre-processing method explained 1.6%, the classifier type explained 1.3% and the feature selection method explained 0.5% (the last three were non-significant after adjustment). These results are in agreement with the conclusions based on the analysis of the data from all teams (see above).

3.5 Modeling practices that lead to best performance

We have conducted several univariate analyses using the modeling factors described in a structured post-challenge survey to which 21 of the participating teams responded. One of the questions asked was whether the team had used cross-validation (on the training dataset) to tune any of the parameters in their classification pipeline. Using a mixed-effect model in which the team and endpoint were treated as random effects and considering the use of cross-validation as a fixed effect, revealed that, on average, the teams that had used cross-validation had *Z*-scores that were higher by 1.2 units for BCM ($P=0.026$), higher by 1.9 units for AUPR ($P=0.023$), whereas for CCEM the scores were not improved significantly. Notably, the partition of training dataset into folds during cross-validation is typically done at random; this can affect the structure of the model (e.g. input features) and its predictions. Although the impact of these stochastic effects on the model performance are typically small, for the psoriasis sub-challenge (where the differences between the 1st and 12th ranked teams were within 2% for all three metrics), the effect of these stochastic effects on the team ranking was substantial (see the Supplementary Note: Impact of random data partitioning in cross-validation on classifier performance).

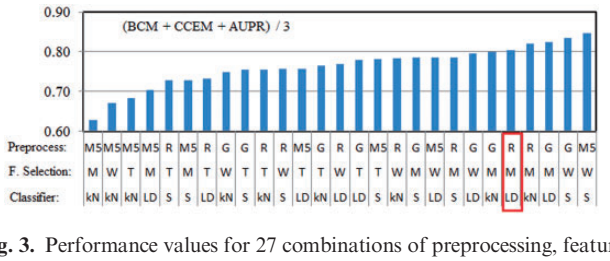

Fig. 3. Performance values for 27 combinations of preprocessing, feature selection and classification methods based on the classification pipeline of the best overall team. Values are average over the four endpoints and the three metrics. Perfect classification is 1.0, whereas 0.0 is perfect misclassification. The combination of methods used by the best overall team is marked in red. T, regular *t*-test; M, moderated *t*-test; W, Wilcoxon rank test; kN, kNN; LD, LDA; S, SVM; M5, MAS5; R, RMA; G, GCRMA

The use of more than one microarray platform in a given challenge did not affect significantly the prediction performance of the classifier. Typically, multiple platforms were used by the teams to increase the number of training samples. Although a larger training dataset is, in general, good for the performance of the classifier, detrimental biases in the process of merging the different platforms can be introduced. These two contradictory effects seem to cancel each other. Tests for the association of other modeling factors with the classifier performance were not feasible, either because of missing data or because of too many factor levels that lead to confounding between the modeling and team factors.

The data generated from the factorial experiment described above and plotted in Figure 3 were also used to identify the modeling methods that seem to work best in general. An ANOVA analysis of the classifier performance metrics as a function of modeling factors for each endpoint separately revealed that the importance of the modeling factors is problem and metric dependent (see Supplementary Fig. S7). Moreover, the best method at each step in the classification pipeline is also problem and metric dependent (see Supplementary Figs. S8–S10). Significant interactions were also present between different factor levels; therefore, using the methods that seem to work best, in general, will not necessarily provide the best possible combination of methods for every dataset. For example, the classification pipeline that maximized the average over all three metrics in all four endpoints used MAS5 preprocessing, gene ranking via Wilcoxon test and an SVM classifier (Fig. 3). However, MAS5 preprocessing was only used twice in the top 10 combinations of methods shown in Figure 3; GCRMA, on the other hand, appeared five times. Moreover, MAS5 was the preprocessing method that appeared most frequently in the worst 10 combinations. So, although GCRMA worked better in combination with most classifier and feature selection methods that were tested than MAS5 did, GCRMA was not part of the best combination of methods. Similarly, the moderated *t*-test appeared six times in the best 10 combinations of methods, yet it was not part of the best single combination. The performances of the LDA and SVM classifiers were similar, whereas kNN was selected only twice in the top 10 combinations and appeared most frequently among the worst 10 combinations. Notably, the combination of methods that was used by the best overall team (RMA, moderated *t*-test and LDA) was ranked fifth of the 27 combinations shown in Figure 3.

Because the training and test datasets were provided at the beginning of the competition, several teams preprocessed the multiple training datasets and the test dataset together using RMA or GCRMA. These two preprocessing methods use

quantile normalization to equalize the probe intensity distributions between arrays. Furthermore, these two preprocessing algorithms use additional parameters that depend on which of the arrays are preprocessed together. We found that, on average, both BCM and AUPR were improved by 6 and 4%, respectively, (Wilcoxon test $P<0.05$) when all the training and test datasets were preprocessed together as opposed to separately (see Supplementary Fig. S11).

3.6 Biomedical factors affecting the quality of predictions

Many biomedical factors can contribute to variability in the classification performance of the models. First, health and disease states, and especially stages of diseases, do not necessarily correspond to well-defined discrete states. This could lead to uncertainties in establishing an exact diagnosis or to less stable signatures of specific diseases/stages. Both the inability to precisely define disease states and inaccuracies in the assessment could have produced artifacts in the phenotype labels of the test samples that we treated as gold standard. Such artifacts would be reflected in the inability of the teams to 'correctly' classify certain samples, as quantified by the misclassification rate across all the teams for those samples.

To identify systematic biases in the classification performance that could be caused by biomedical factors or a questionable gold standard, we correlated the misclassification rate across all the teams with metadata that was not originally disclosed to the participants. The metadata comprised disease characterization, demographic factors and, in the LC challenge, an independent reevaluation of the histology of the tumors. In psoriasis, disease onset is gradual, and hence the disease status of a patient may be ill defined at the earliest stages of the disease. Figure 4a shows that the highest misclassification rate for a psoriasis sample was found for a patient with a Psoriasis Area and Severity Index (PASI) score (Langley and Ellis, 2004) of 3.2 (low severity); samples with the highest PASI scores were rarely misclassified. The Spearman rank correlation between the misclassification rate and PASI scores was −0.53 ($P=0.011$). In the LC sub-challenge, samples with an inconclusive or divergent second diagnosis had significantly higher average misclassification rates (38 and 56%, respectively, $P<0.01$) compared with the average misclassification rate (19%) for samples with concordant second diagnoses (Fig. 4b).

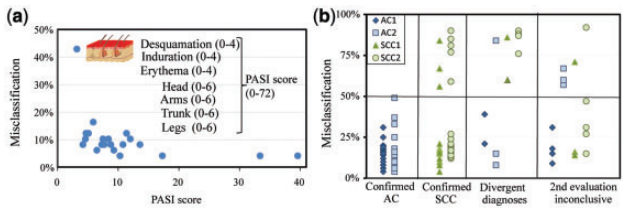

Fig. 4. (**a**) Misclassification rate of psoriasis samples as a function of the PASI score. (**b**) Misclassification rates in the LC sub-challenge shown as a function of the agreement between two independent pathological evaluations (confirmed, divergent and inconclusive diagnoses in the second evaluation). AC, adenocarcinoma; SCC, squamous cell carcinoma; post-fixes 1 and 2 indicate the cancer stages

3.7 Persistent genes in disease signatures across multiple teams

Using the data from the post-challenge survey, we performed a meta-analysis to identify the genes that were used by more than one of the participating teams as inputs in their prediction models. Out of submitted gene lists (five for MS Diagnostic, nine for psoriasis, six for COPD and seven for LC), we retained only the gene lists of the teams whose models performed better than chance ($P<0.05$) for at least one of the three metrics (two for MS Diagnostic, nine for psoriasis, five for COPD and seven for LC). The number of times that any one gene occurred across the teams (gene frequency) was determined, and the observed average gene frequency was computed for each endpoint. A simulation analysis was conducted to determine if the average gene frequency was greater than expected by chance by randomly generating lists of genes (the same number and size as the observed list) from all the genes measured on the microarrays. For both LC and psoriasis, the overlap of gene lists between the teams was greater than expected by chance ($P<0.0001$). The number of genes submitted by two or more of the teams was, none for MS Diagnostic, 2 for COPD (namely CDKN1C and CDKN2A), 247 for LC and 216 for psoriasis. The most frequently occurring genes in the signatures of the different teams for psoriasis and LC are given in Supplementary Tables S3 and S4, respectively.

4 DISCUSSION

With 54 teams participating from around the world, the IMPROVER DSC provided an important opportunity to test the feasibility of crowdsourcing to answer computational biology questions relevant to both industry and academia. We found that most of the 54 different teams used a similar classification pipeline for all the sub-challenges, with adjustments as required by the specificity of each endpoint. The differences in the quality of predictions between teams for any endpoint were mainly the result of three factors: (i) the training datasets used, (ii) the underlying classification pipelines and (iii) the skill level of the team in applying/tuning the pipeline to each endpoint. The effect of the skill factor was highlighted previously by the MicroArray Quality Control (MAQC)-II consortium (Shi *et al.*, 2010), a large scale initiative completed over a few years that also studied the feasibility of microarray-based predictions and classifier development practices.

Although there were significant differences in the performance between teams, and the overall team ranking for the top teams was stable, the differences in performance were not enough to allow us to identify one classification pipeline as the best practice in all cases. For example, the best overall team was not the best performer in any of the individual sub-challenges. In fact, *no single team performed best in more than one sub-challenge/endpoint*. The likely main reason for this finding is that the *best method to use at each step in the classification pipeline is endpoint dependent*, whereas most teams used a more or less consistent pipeline for all sub-challenges. In general, it is to be expected that a one-size-fits-all approach would not be superior to a more tailored approach on all datasets. Even if each team had used the same training datasets and searched for the best method at each step in the classification pipeline, it is still unlikely that the same classification pipeline/

team would have emerged at the top of the rankings in most or all the four sub-challenges due to stochastic factors present in most pipelines. For example, for *Team 221,* variations because of the random way in which cross-validation partitions were generated during the tuning of the classification pipeline could have produced an increase of 11 rank units in the psoriasis sub-challenge, assuming that everything else in the pipeline was kept the same (see Supplementary Note: Impact of random data partitioning in cross-validation on classifier performance).

The classification pipelines of the best overall three teams had some common and some different elements. For example, the first and third teams preprocessed the training and test datasets together using RMA, whereas the second and third teams preselected features using a Wilcoxon test. Looking at the classification models used, the first team used a classical discrimination method (LDA), the third team used LASSO logistic regression (an emerging approach to high-dimensional data classification) and the second team used an experimental clustering-based classifier. Both the first and third teams tuned the number of features to be used in each particular problem using cross-validation, whereas the second team used trial and error to determine the number of features. The number of features that the teams used did not correlate with the ranking of the top three teams; the first team used only two genes in two of the four sub-challenges and, at most, 25 genes in the other two, the second team used hundreds of genes in all four sub-challenges, whereas the third team used between 14 and 60 genes.

The observed variability in the performance of the models (as an average over the three metrics) was mainly (70%) explained by the endpoint, which is similar to the results reported by the MAQC-II initiative (65%) using the MCC metric. The team/classification pipeline explained only about 8% of the variance. Similar results were obtained from a factorial experiment designed to quantify the contribution of the preprocessing method, feature selection method and classifier type in addition to the endpoint. The endpoint was again found to be the only significant contributor to the total variance in the performance data, explaining 65% of the variance. When the variance in performance data was examined for each endpoint separately (Supplementary Fig. S7), we concluded that *no single modeling factor (e.g. the classifier type) is always the biggest contributor to the performance of the pipeline*, and hence *optimization of the choice of method at each stage in the classification pipeline is needed for optimal result*. Such multifactorial optimization is time-consuming and, for simplicity, most modelers prefer to fix some of the modeling factors to their method of choice while trying to optimize others. To facilitate the development of microarray-based prediction models by the community at large, we made available a software package called *maPredictDSC* that implements the combination of methods used by the best overall team in the IMPROVER DSC. In addition, the package offers the possibility to automatically search among a large number of combinations of methods for the one that maximizes the cross-validated performance on the training data. An implementation of the 'wisdom of crowds' is offered by combining the predictions of the resulting models. Our analyses of the DSC results suggest that the best practices that could work best in general are as follows:

(1) GCRMA seemed to work better than MAS5 in most situations, yet the limitations of MAS5 could be mitigated, at least in part, by using existing batch correction methods that were not considered in this study but which have been discussed previously (Luo *et al.*, 2010).

(2) RMA or GCRMA preprocessing of all training and test datasets together was better than preprocessing them separately. This type of preprocessing was used by the first and third best overall teams, and the results from a controlled experiment supported this conclusion. Combined preprocessing of the datasets was possible in the IMPROVER DSC but not in MAQC-II, because the test expression data were made available to the participating teams at the beginning of the DSC. Although it could be argued that in a clinical setup, only one test sample gene expression profile would be available, and hence a full test dataset would not be available to preprocess with the training datasets; the merit of preprocessing multiple training datasets together versus separate preprocessing (with or without applying other batch correction methods) should be investigated further.

(3) Typically, the ordinary *t*-test underperformed when compared with a moderated *t*-test or Wilcoxon test for feature ranking.

(4) The use of cross-validation in optimizing classification pipeline parameters was important. The first and third best overall teams used such an approach, and mixed-effects modeling of the performance data from a population of models also supported this conclusion.

(5) Although the LDA classifier was used by the best overall team in all the sub-challenges, the results from the factorial experiment analysis on the same pipeline showed that a similar performance could have been obtained with SVM instead of LDA. The poorer performance observed for the kNN classifier can be explained, at least in part, by the fact that the value of $k = 3$ in the kNN was chosen arbitrarily and not optimized for each sub-challenge.

In addition to the conclusions that can be drawn from the outcome of the DSC from a computational perspective, our analysis of the overall misclassification rate for each test sample across all teams highlighted the persistent challenges in transcriptomics-based disease classification. Binary classification of clinical samples, e.g. in the psoriasis sub-challenge, cannot take into account the facts that disease progression is likely more continuous than discrete, and that there may be no clear border between disease classes from a clinical and a molecular perspective. In the LC sub-challenge, the gold standard may have been imperfect because established diagnosis techniques (e.g. medical examination and histology) are not 100% reliable. To investigate the robustness of the gold standard, the histological images from the LC specimens were reevaluated by a second pathologist, and we have found that on average, for samples where the two examiners did agree, the misclassification rate was lower than for the remaining instances.

The participation of many different teams in the IMPROVER DSC allowed us to examine the methods and practices currently in use in computational biology. Even though the teams worked independently and competed for a prize ($50 000 to be used for research purposes), they were also collaborating in testing the

prediction methods. Overall, the competition was transformed into a collaboration that will benefit the community at large. Because the training datasets in the IMPROVER DSC were composed of multiple heterogeneous public datasets *totally unrelated* to the test datasets, except for studying the same phenotype, this challenge was a stringent and unprecedented test of the ability of models to predict outcomes from gene expression data that better reflected their likely clinical applications.

ACKNOWLEDGEMENTS

We thank Filipe Bonjour, Bruce O'Neel and Kaushik Sarkar for the development and support of the DSC Web site; Claudia Frei, Christian Haettenschwiler and Shweta Stadler for the organization of communications with the DSC participants and of the IMPROVER symposium 2012; Jean Binder, Elise Blaese, Nathalie Chemineau, Stephanie Corthesy, Peter Curle Fraser, Chaturika Jayadewa, Lionel Schilli, Joerg Sprengel and all the team members, who are not among the authors, for their contributions during discussion sessions and for the management of the project.

A full list of IMPROVER DSC collaborators is included as Supplementary Material

Funding: Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of Health and Human Services (N01-HD-2-3342) (A.T. and R.R., in part). DSC best overall performer grant from Philip Morris International (to A.T.). Swiss National Science Foundation (PP00P2_128503) and from SystemsX.ch, the Swiss Initiative for Systems Biology (to M.U., P.N., C.Z. and H.K.).

Conflict of Interest: none declared.

REFERENCES

(2006) NSABP study confirms oncotype DX predicts chemotherapy benefit in breast cancer patients. *Oncology (Williston Park)*, **20**, 789–790.

Acharya,C.R. *et al.* (2012) Retraction: Acharya CR, *et al.* Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. JAMA. 2008; 299:1574–1587. *JAMA*, **307**, 453.

Baldi,P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, **57**, 289–300.

Bittner,M. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Gordon,G.J. *et al.* (2002) Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, **62**, 4963–4967.

Hresko,A. and Haga,S.B. (2012) Insurance Coverage Policies for Personalized Medicine. *J. Pers. Med.*, **2**, 201–216.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kaklamani,V. (2006) A genetic signature can predict prognosis and response to therapy in breast cancer: Oncotype DX. *Expert. Rev. Mol. Diagn.*, **6**, 803–809.

Langley,R.G. and Ellis,C.N. (2004) Evaluating psoriasis with Psoriasis Area and Severity Index, Psoriasis Global Assessment, and Lattice System Physician's Global Assessment. *J. Am. Acad. Dermatol.*, **51**, 563–569.

Lauria,M. (2013) Rank-based transcriptional signatures: a novel approach to diagnostic biomarker definition and analysis. *Syst. Biomed.*, in press.

Luo,J. *et al.* (2010) A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenomics J.*, **10**, 278–291.

Marbach,D. *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Meyer,P. *et al.* (2011) Verification of systems biology research in the age of collaborative competition. *Nat. Biotechnol.*, **29**, 811–815.

Meyer,P. *et al.* (2012) Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. *Bioinformatics*, **28**, 1193–1201.

Mook,S. *et al.* (2007) Individualization of therapy using MammaPrint: from development to the MINDACT Trial. *Cancer Genomics Proteomics*, **4**, 147–155.

Nandy,P. *et al.* (2013) Learning diagnostic signatures from microarray data using L1-regularized logistic regression. *Syst. Biomed.*, in press.

Nguyen,B. *et al.* (2012) Comparison of molecular subtyping with BluePrint, MammaPrint, and TargetPrint to local clinical subtyping in breast cancer patients. *Ann. Surg. Oncol.*, **19**, 3257–3263.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.

Schena,M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shi,L. *et al.* (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Tarca,A.L. *et al.* (2007) Machine learning and its applications to biology. *PLoS. Comput. Biol.*, **3**, e116.

Tarca,A.L., Than,N.G. and Romero,R. (2013) Methodological Approach from the Best Overall Team in the IMPROVER Diagnostic Signature Challenge. *Syst. Biomed.*, in press.

Tibshirani,R. *et al.* (1996) Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58**, 267–288.

Yamani,M.H. *et al.* (2007a) Post-transplant ischemic injury is associated with up-regulated AlloMap gene expression. *Clin. Transplant.*, **21**, 523–525.

Yamani,M.H. *et al.* (2007b) Transplant vasculopathy is associated with increased AlloMap gene expression score. *J. Heart Lung Transplant.*, **26**, 403–406.