IMPROVER
systemsbiologyverification

Computational Lessons

Gustavo Stolovitzky, Raquel Norel,
Erhan Bilal, IBM Research

3rd October 2012

On the shoulder of giants: The MAQC-II study on validation of microarray-based predictive models

- MAQC-II is an unprecedented collaborative research project spearheaded by the FDA with participation of other government agencies, industry and academia.

- The MAQC-II study aimed to establish guidelines for microarray-based predictive models by
  - evaluating sources of bias in study design, taking into account population heterogeneity,
  - surveying approaches in genomic model development,
  - understanding the sources of model performance variability and
  - assessing the influence in model performance of end point signal strength

- The IMPROVER Diagnostic Signature Challenge (DSC) aimed to establish that it is possible to blindly predict phenotype from transcriptomics data as an example of industrial verification of a research building block.

- Despite the difference in motivation between IMPROVER and MAQC-II, the methods employed by the two studies are very similar, and indeed our conclusions closely follow those of MAQC-II: This is a good example of reproducibility in science!

Comparison between MAQC-II and IMPROVER

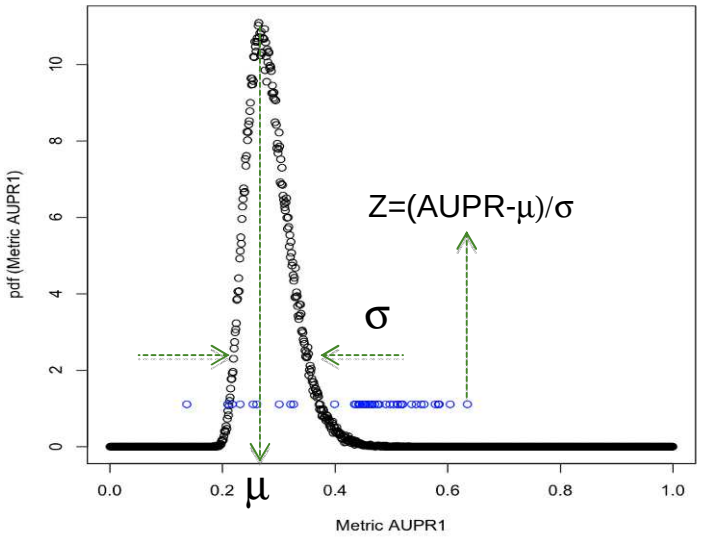
	MAQC-II project	IMPROVER project
Project Type	Collaboration (includes review of method by other members)	Competition
Submission	predictions (+ model summary information tables)	predictions (+ write up)
Number participating teams	36 17 submitted to all endpoints	55 33 submitted to all endpoints
Classification type	Binary	Confidence levels: 2-way + 4-way
Number test datasets	6	5
Species	human, mouse, rat	Human
Diseases	Lung & breast cancer, Multiple Myeloma, Neuroblastoma	Psoriasis, COPD, lung cancer, Multiple Sclerosis
Prediction type	13 preclinical and clinical endpoints, 2 positive controls and 2 negative controls	Diagnosis, staging
Training Datasets	Public	Public
Test Datasets	Made for MAQC-II project by the same data providers as the training data set. Biological endpoints unknown to modelers	Created for or licensed to the IMPROVER project. Completely unrelated to the training datasets
Project stages	Stages 1: train on training set and test on blind test dataset Stage 2: train on test and test on training dataset (not blind)	Only one stage: train on training dataset, test on test dataset
Scoring metrics	• Matthews Correlation Coefficient (MCC) • Accuracy • Sensitivity • Specificity • Area under the receiver operating characteristic curve Root mean squared error	• BCM (Belief Confusion Metric) • CCEM (Correct Class Enrichment Metric) • AUPR_avg (Area Under the Precision-Recall score)
Methods used	• 17 summary and normalization methods • 9 batch effect removal methods • 33 feature selection methods • 24 classification algorithms	• MAS5, RMA, fRMA, GCRMA, Hook, Combinations • More than 15 feature selection methods • More than 10 classification algorithms
Gain for community	• Available datasets can be used for benchmarking • Series of publication summarizing key lessons and biological interpretation	• Available datasets can be used for benchmarking • Determine the existence of a robust signature for a particular disease/data set • Methods to be published in special issue of Systems Biomedicine

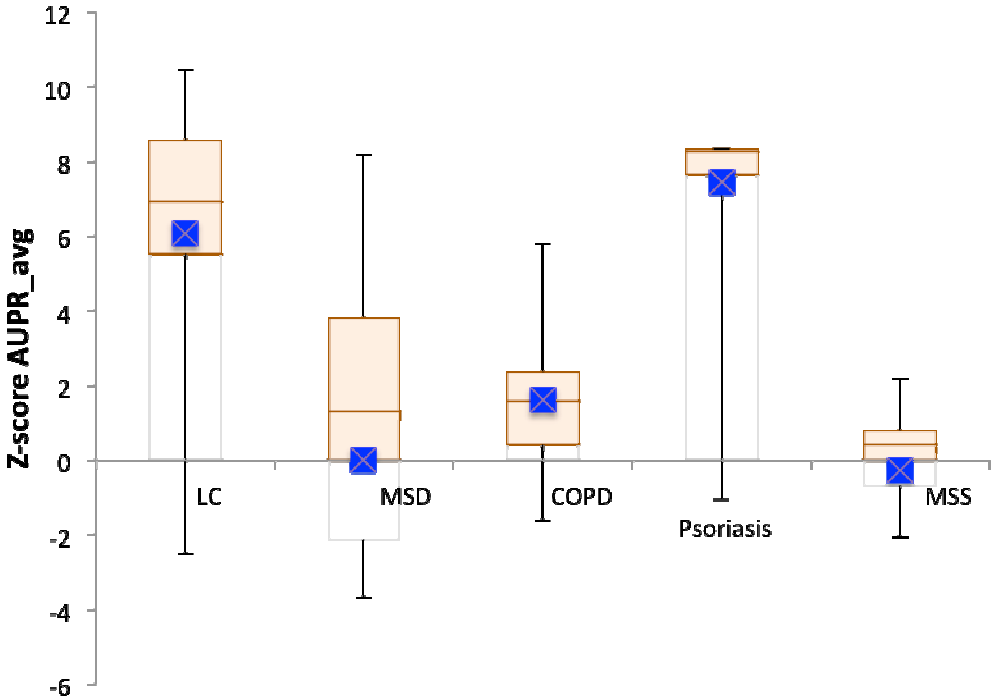
Comparison between MAQC-II and IMPROVER

	MAQC-II project	IMPROVER project
Project Type	Collaboration (includes review of method by other members)	Competition
Classification type	Binary	Confidence levels: 2-way + 4-way
Test Datasets	Made for MAQC-II project by the same data providers as the training data set. Biological endpoints unknown to modelers	Created for or licensed to the IMPROVER project. Completely unrelated to the training datasets

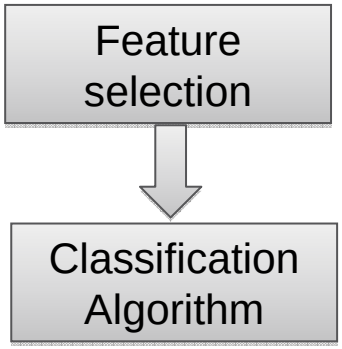
Main computational lessons from IMPROVER

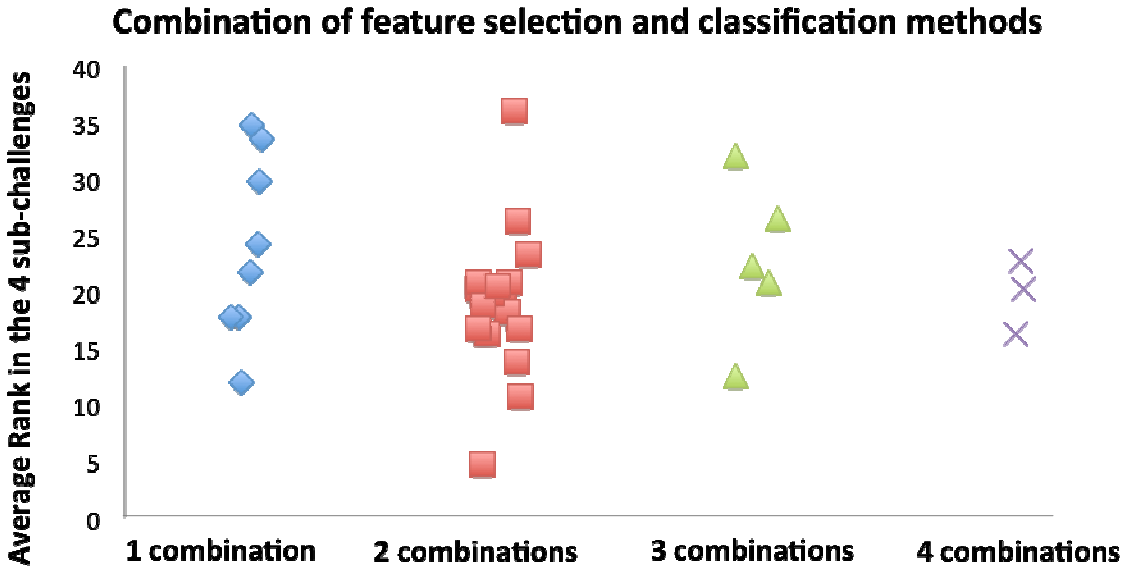
- The degree of difficulty of a sub-challenge (given, e.g., by the median AUPR z-score) can be used to quantify the endpoint signal strength in a given classification problem.
- Transcriptomics data from Psoriasis is the most informative, followed by LC, COPD and MSD. MSS has little if any signal.





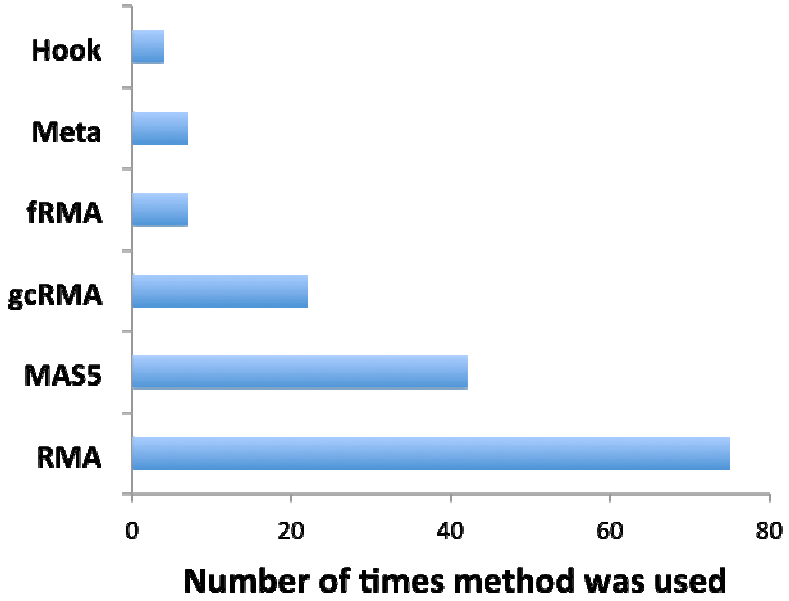
Tailoring the analysis pipeline to a particular dataset, on its own, did not improve overall performance



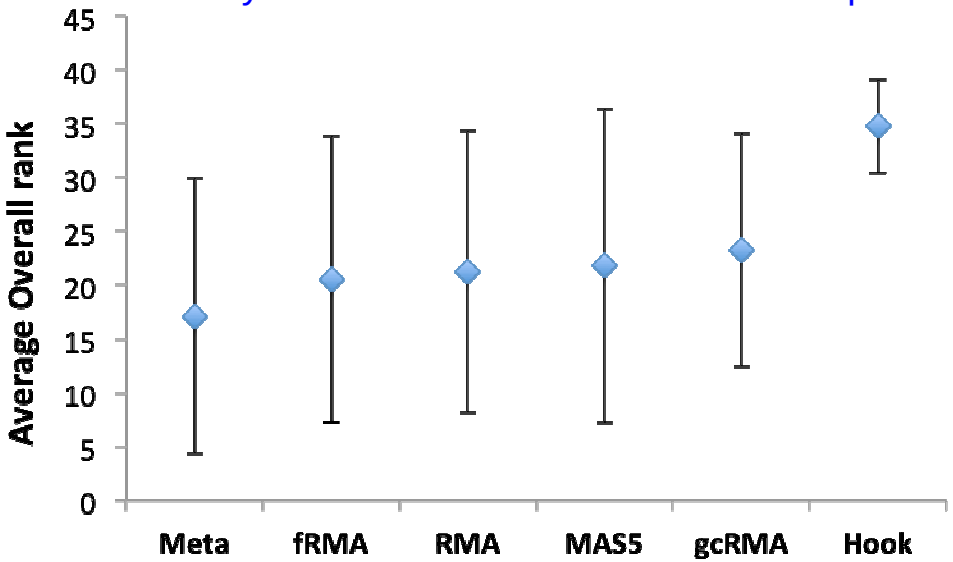


- The degree of difficulty of a sub-challenge could have been detected in training (MAQC-II)
- Some teams made different combinations of feature selection and classification algorithms to deal with the subtleties of each sub-challenge.
- Tailoring the classification pipeline, however, did not confer a clear advantage.

A variety of normalization methods was used across all sub-challenges

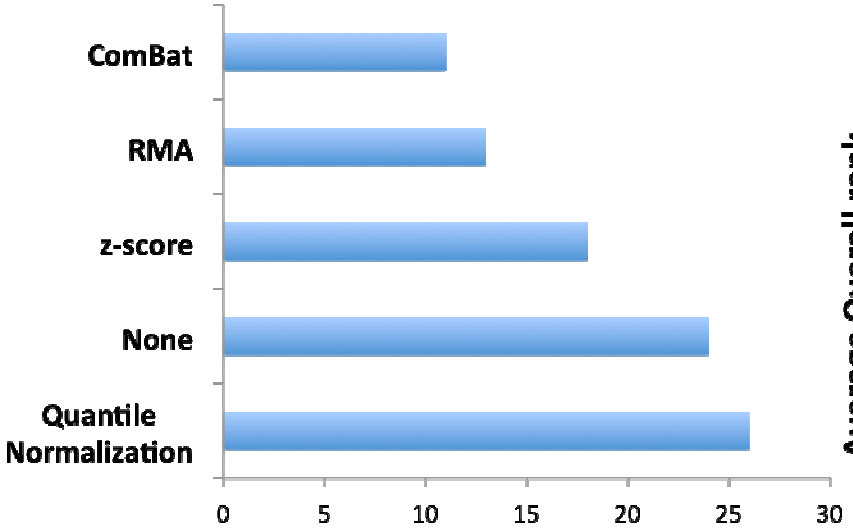


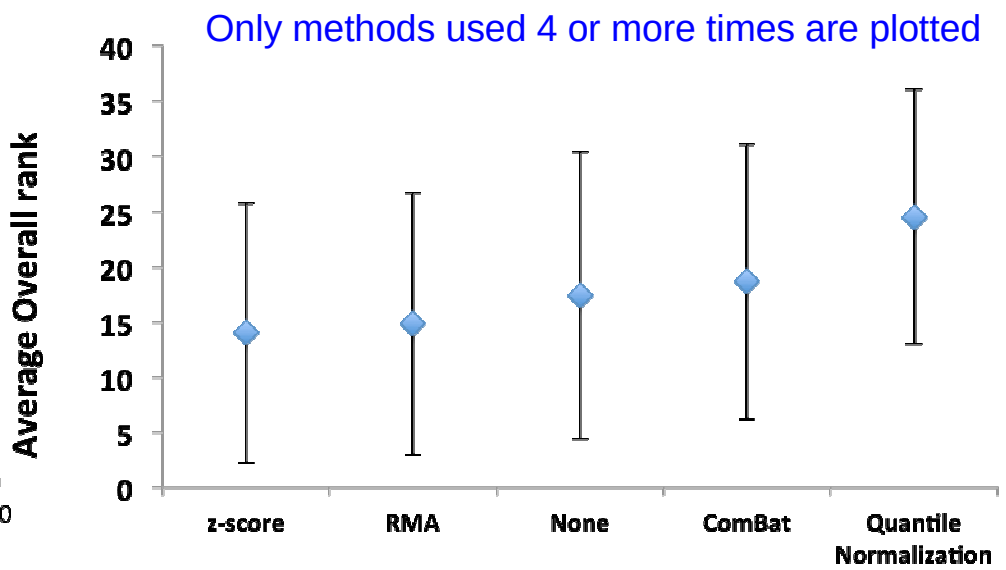
Only methods used 4 or more times are plotted



- RMA was the normalization method most used across all sub-challenges
- Teams that normalized using Hook, a method used in 4 sub-challenges, attained the worst overall rank (but statistics N=4 is very small)
- No normalization method seems to have been key for attaining a good ranking.

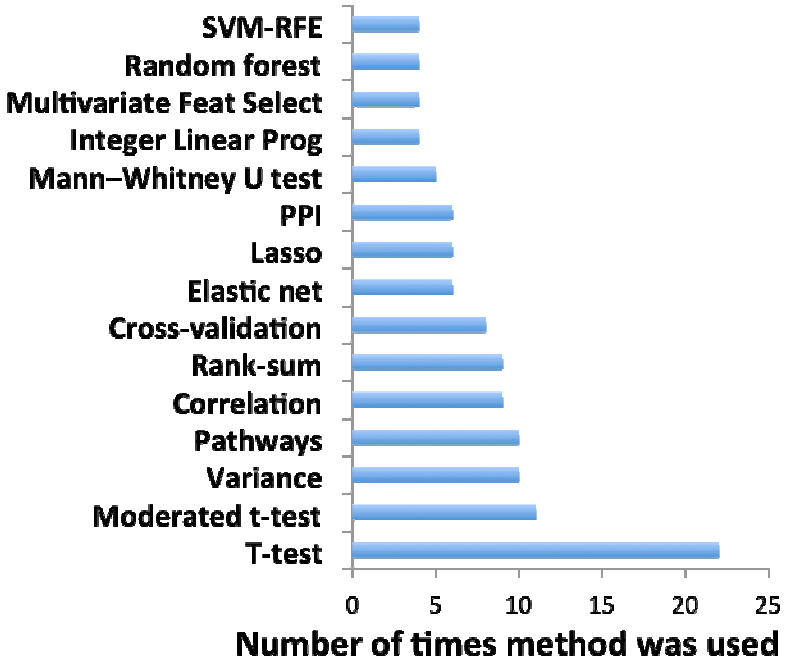
A variety of Batch Effect Correction methods was used across all sub-challenges

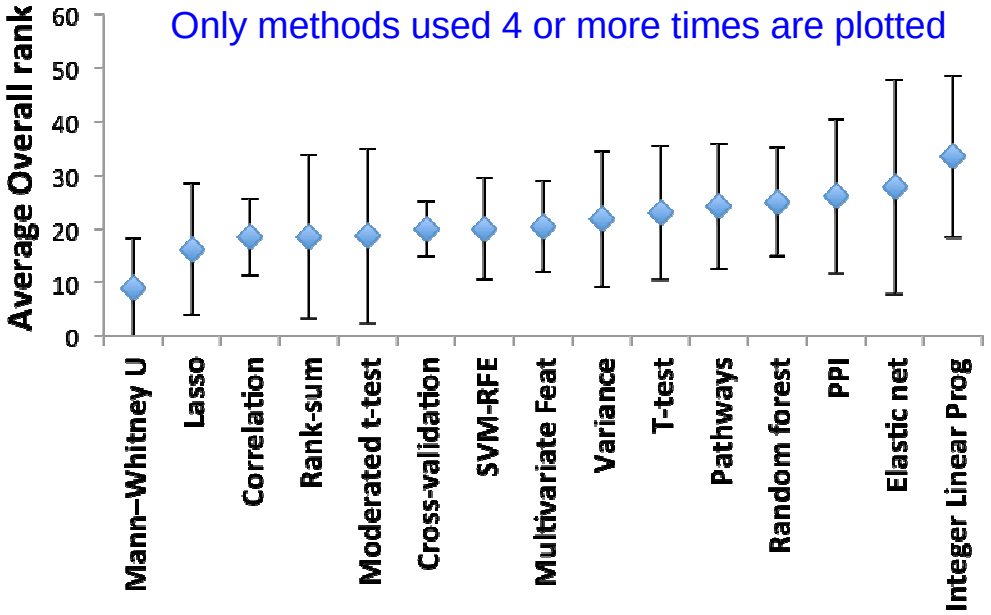




- Quantile Normalization was the Batch effect correction method most used across all sub-challenges (26 times).
- In 24 instances, no batch effect correction method was used.
- No batch effect correction method seems to have been key for attaining a good ranking.
- Other batch effect correction method used were: Empirical Bayes, removeBatchEffect from limma, Houskeeping genes, Linear regression, GCRMA, SVA and XPN.
- XPN got an average overall rank of 6.67, but the statistics is scarce (N=3).

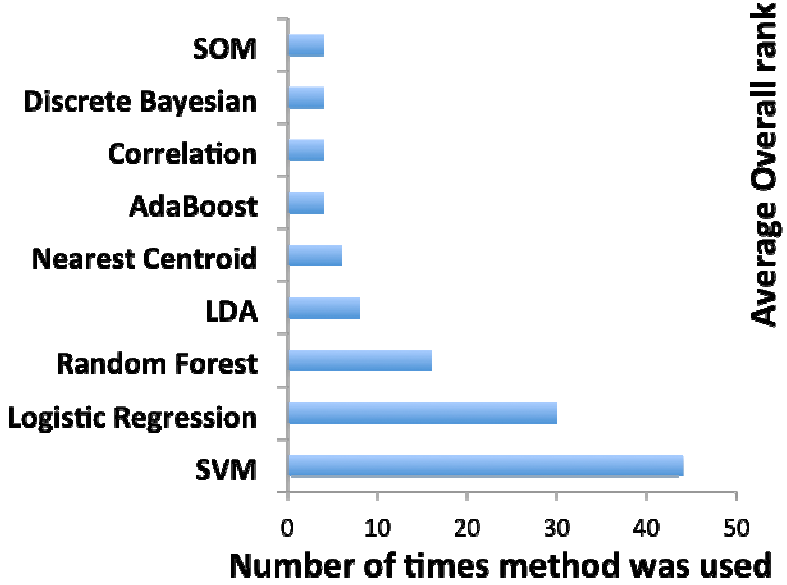
A variety of feature selection methods was used across all sub-challenges



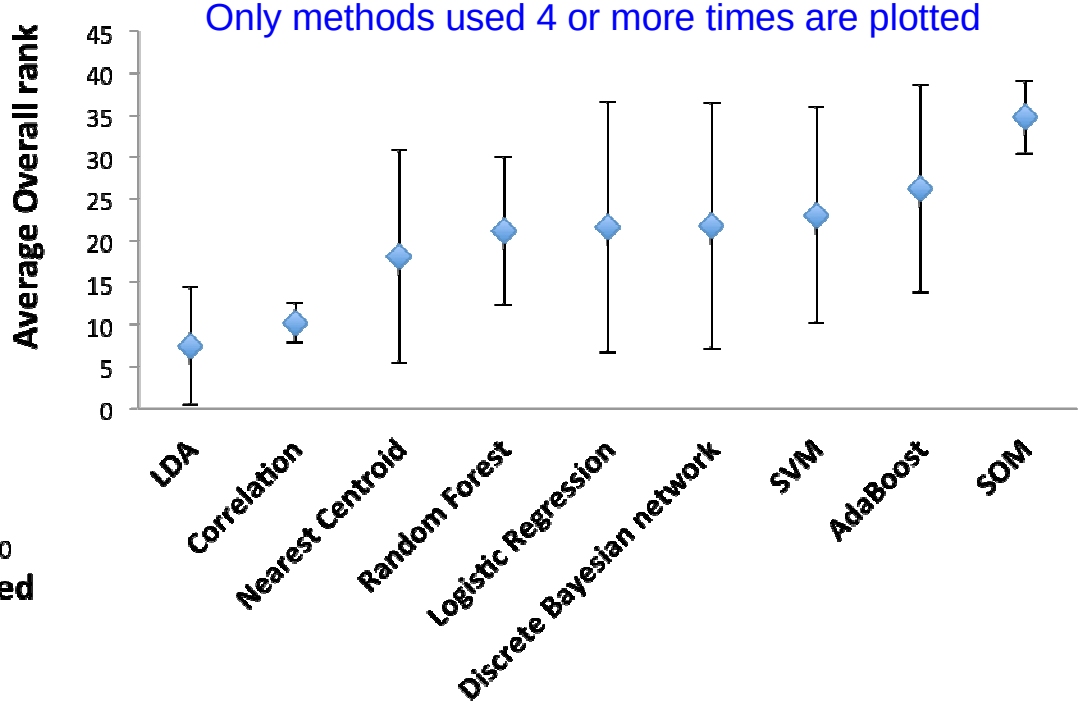


- T-test and moderated t-test were the feature selection methods most used across all sub-challenges
- Teams that used Mann-Whitney U test, a method used in 5 sub-challenges, attained on average the best overall rank of 8.9, followed by Lasso with a rank of 16 (but N=5 is small)

A variety of classification algorithms were used across all sub-challenges



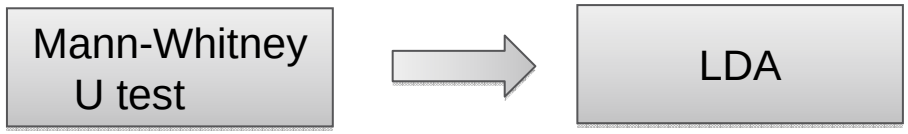
Only methods used 4 or more times are plotted



- SVM and logistic regression were the classification methods most used across all sub-challenges
- LDA was used 8 times. On average, LDA attained the best overall rank of 7.5, followed by correlation methods with a rank of 10.25 (but N=4 for the latter)
- Interestingly, LDA was the classification method with the largest positive impact in the final MAQC-II scores.

A winning combination?

- The previous univariate analysis of methods suggest a few good combinations for the task of feature selection and classification:



- This combination was used only once, by the best performing team in Lung Cancer. That team did not submit the other sub-challenges.

Methods used by the best performers

IMPROVER systemsbiologyverification

Team	Normalization	Batch effect	Feature selection	Classification algorithm
36	RMA	None	**Mann-Whitney U test**	**LDA**
221	MAS5	Quantile normalization	**Moderated t-test**	**LDA**
114	fRMA	ComBat	**Moderated t-test**	Threshold Gradient Descent Regularization

Lung Cancer

Team	Normalization	Batch effect	Feature selection	Classification algorithm
227	MAS5	None	**Mann–Whitney U test**	Nearest centroid
208	RMA	z-score	Elastic net	**Logistic regression**
221	RMA	Quantile normalization	**Moderated t-test**	**LDA**

MSD

Team	Normalization	Batch effect	Feature selection	Classification algorithm
294	RMA	RMA, housekeeping genes	**Moderated t-test**, Scaled alignment selection	Kernel Fisher discriminant
161	RMA	RMA	Rank-sum, Lasso	**Logistic regression**
50	MAS5	None	Variance, IQR, linear fit, ROC **Moderated t-test**	Neural network

Psoriasis

Team	Normalization	Batch effect	Feature selection	Classification algorithm
122	MAS5, quantile normalization	Combat		random generalized linear model predictor (RGLM)
221	RMA, quantile normalization		differential expression statistics	**LDA**
112	RMA	zero mean and unit variance, quantile normalization	T-test	**Logistic Regression**

COPD

LDA, Mann-Whitney, Moderated T-test and logistic regression: Was it a fluke?

- LDA appeared 4 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.006 (Binomial test, N=12, k=4, p=8/120). **Significant**

- Logistic regression appeared 3 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.6 (Binomial test, N=12, k=3, p=30/120). **Not Significant**

- Mann-Whitney U test appeared 2 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.06 (Binomial test, N=12, k=2, p=5/151). **Marginally Significant**

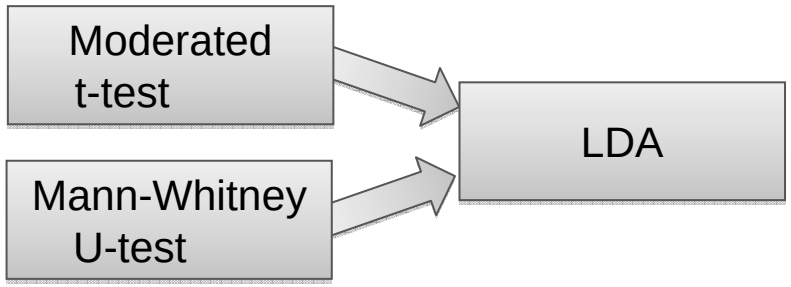
- Moderated t-test appeared 5 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.001 (Binomial test, N=12, k=5, p=11/151). **Significant**

- Moderated t-test in combination with LDA appeared 2 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.002 (Binomial test, N=12, k=2, p=11*8/(120*151)). **Significant**

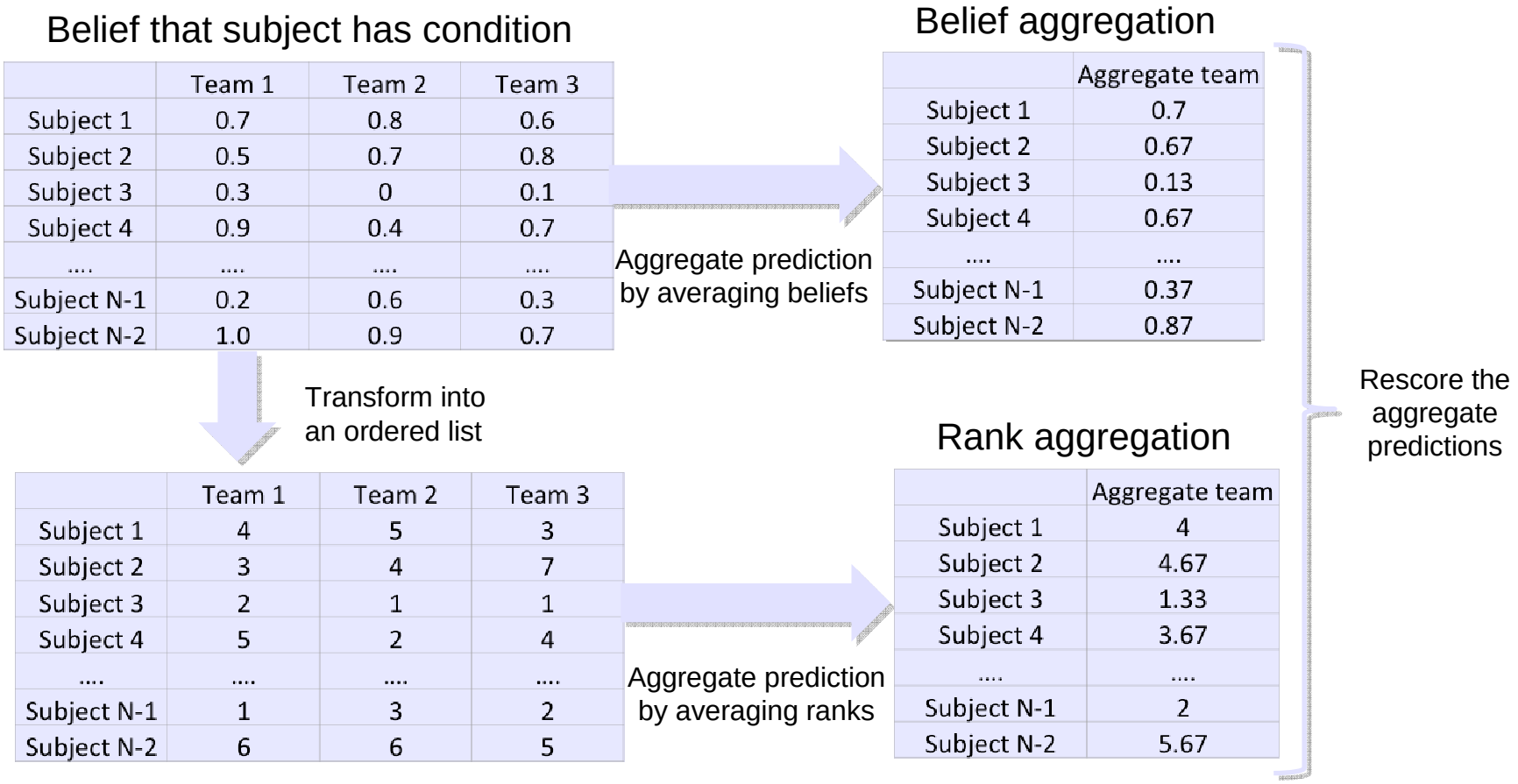
- Mann-Whitney U test in combination with LDA appeared 1 times in the first 3 positions of the 4 sub-challenges. The p-value of that occurrence was 0.025 (Binomial test, N=12, k=1, p=5*8/(151*120)). **Significant**

Team 221 (Overall Best performer) used this combination

Team 36 (Lung Cancer Best performer) used this combination



The Wisdom of Crowds for Diagnostics: aggregating predictions



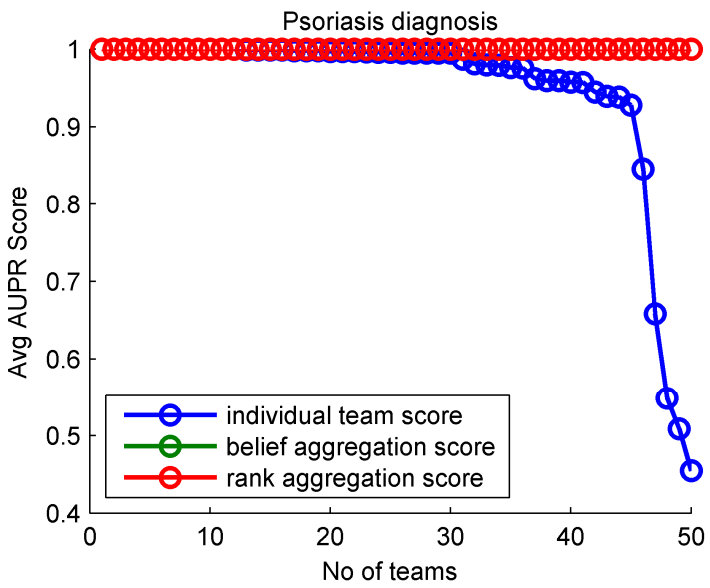
	Team 1	Team 2	Team 3
Subject 1	0.7	0.8	0.6
Subject 2	0.5	0.7	0.8
Subject 3	0.3	0	0.1
Subject 4	0.9	0.4	0.7
….	….	….	….
Subject N-1	0.2	0.6	0.3
Subject N-2	1.0	0.9	0.7

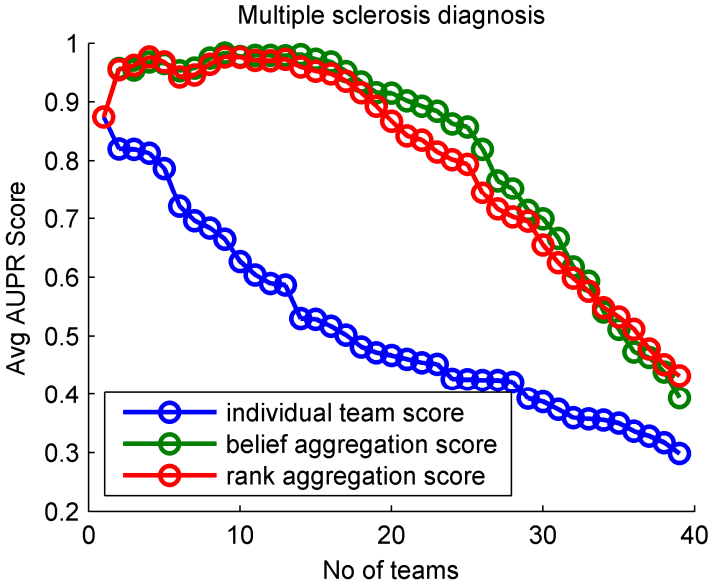
	Aggregate team
Subject 1	0.7
Subject 2	0.67
Subject 3	0.13
Subject 4	0.67
….	….
Subject N-1	0.37
Subject N-2	0.87

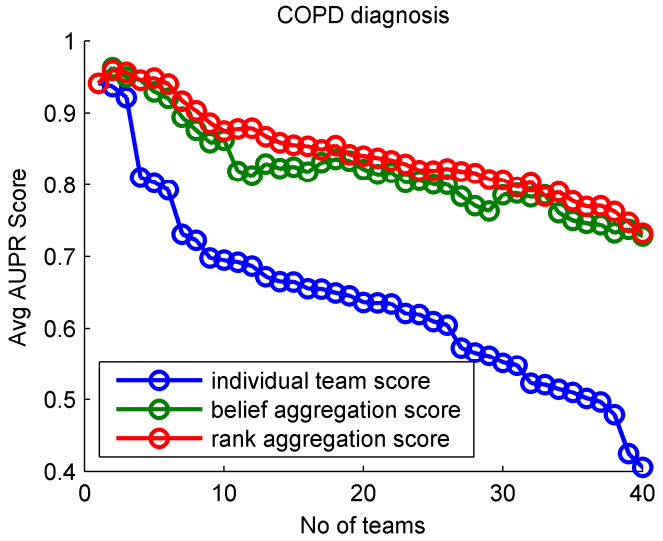
	Team 1	Team 2	Team 3
Subject 1	4	5	3
Subject 2	3	4	7
Subject 3	2	1	1
Subject 4	5	2	4
….	….	….	….
Subject N-1	1	3	2
Subject N-2	6	6	5

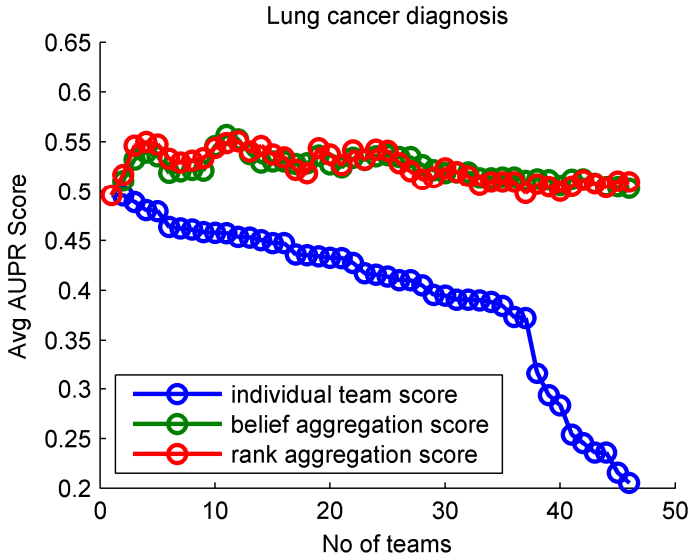
	Aggregate team
Subject 1	4
Subject 2	4.67
Subject 3	1.33
Subject 4	3.67
….	….
Subject N-1	2
Subject N-2	5.67

Wisdom of Crowds at work in IMPROVER

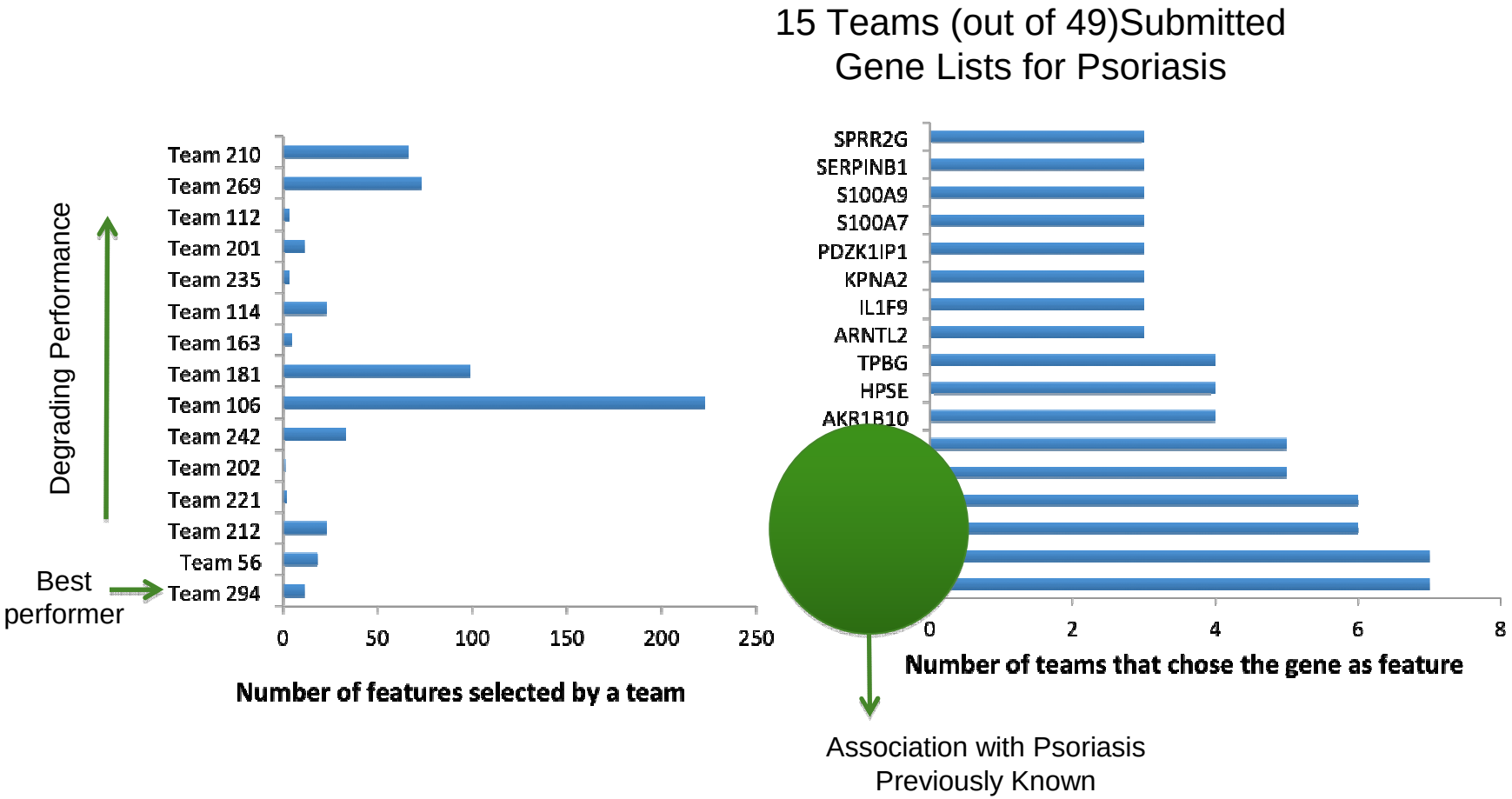








Common Features in Psoriasis



- Team 221 submitted only 2 genes (ranked 12th, with AUPR or 0.9999).
- We are in the process of doing similar analysis for the other diseases.

Lessons and Conclusions (1 of 2)

- Can computational methods perform disease classification from transcriptomics data?
  - The answer is that it depends on the type of disease more than on the computational method.
  - If the phenotype information in the data is faint (due to cell type, data modality or confounding factors), the computational method could play a crucial role.
  - The challenge results themselves provide a quantitative measure of how much signal there is in the data

- Design of challenge data has to avoid confounding batch effects with phenotype effects.

- It may be wise not to give all the data on the test set, as it can provide unintended information to the participants and it better represents the situation at the clinic

- Similar computational methods can have a wide range of performance within the same challenge: no single method resulted as the clear winner
  - However, two feature selection methods (Mann-Whitney U test and Moderated t-test) and a classification algorithm (Linear discriminant analysis) consistently ranked within the top 3 performers.
  - Aggregation of results performed better than the best in 3 out of 4 sub-challenges.

Lessons and Conclusions (2 of 2)

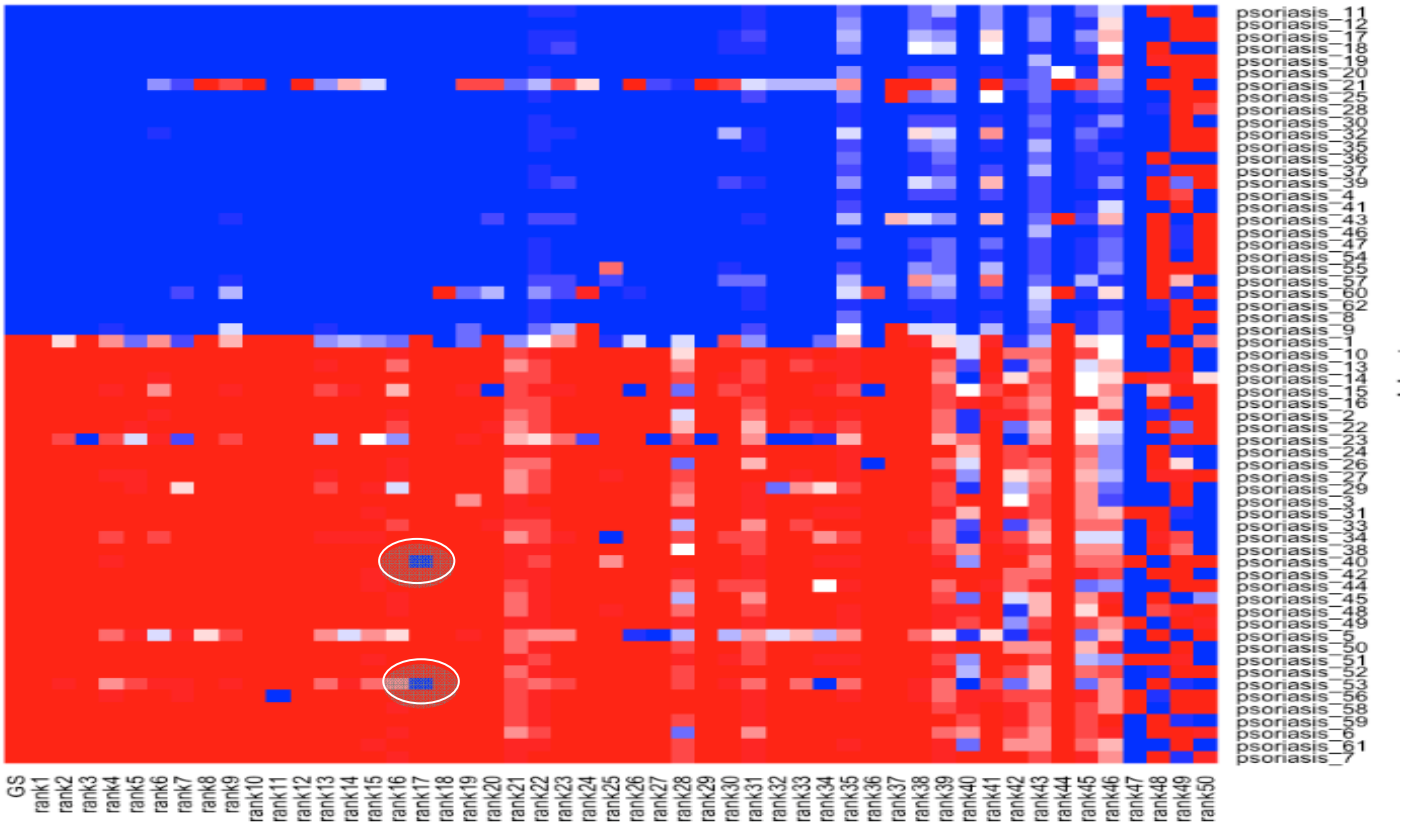
- The challenge was a success in terms of participation, with 55 teams submitting to at least one challenge
  - The advantage of having many participants, however, created a problem with the multiple testing corrections, that diminishes the statistical significance of the results. (MSS)

- Many of these lessons learned in this IMPROVER challenge are consistent and strengthen the conclusions reached in the MAQC-II study (2010).

A Few IMPROVER Anecdotes to Start the Discussion

An IMPROVER story: A One-gene signature

- Teams 202 (ranked 17 with AUPR of 0.95 in Psoriasis) submitted only 1 gene as feature: IGFL1 (insulin growth factor-like family member).
- Only one paper relates IGFL1 and Psoriasis in the literature (PMID: 21454693).
- Yet, team 202 has only 2 misclassified subjects:



Another IMPROVER story: random predictions

- Teams that did not submit a prediction to a challenge, were still scored towards the overall performance by receiving a rank of 54 (worst possible ranking)
- A team decided that rather than not submitting to some of the challenges, they would submit just a random prediction.
- Indeed the write-up of this participant (for the challenges that they didn't participate in) read: "Random Submission".

Team	LC rank	COPD rank	MSD rank	Psoriasis rank	Final Rank
TeamXYZ actual random submission	~70%	~45%	~35%	~25%	~50%
TeamXYZ had they not submitted	100%	100%	100%	~25%	~86%

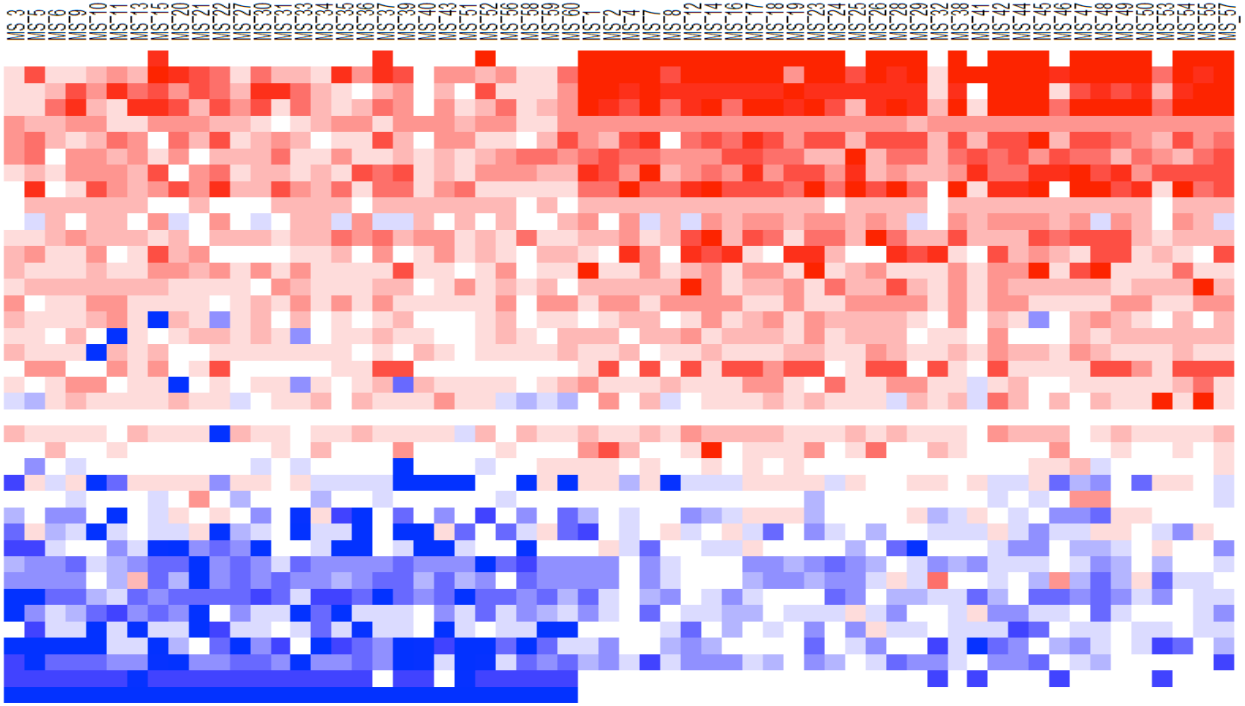
In green are random predictions (with the numbers perturbed a little to anonymize the team)
Clearly, submitting a random prediction beats the penalty of not submitting!

Another IMPROVER story: all-affected or otherwise trivial predictions

In a similar strategy to submit an entry, a team submitted a prediction in which all the labels predicted the subject as affected.

This team ranked in the top 40% in the MSD challenge, which is better than average.

Trivial Submission →



Item for discussion: Is it fair to accept such submissions given that a trivial submission provide no service to the community when even a failed real method provides important information?

Another IMPROVER story: an example of the “self assessment trap”?

One of the submitting teams used a method to predict the class labels of unknown cancer samples, which they had published previous to the IMPROVER challenge.

In their paper, they compare their algorithm and show that theirs outperforms other existing methods for the dataset they trained on.

One of the functions of IMPROVER is to verify such claims of performance, and avoid falling into the self assessment trap.

In the blind prediction experiments, this method did not outperform the others as originally advertised in their paper:

Team	LC rank	COPD rank	MSD rank	Psoriasis rank
TeamABC	~65%	~20%	~20%	~60%

The numbers have been perturbed to anonymize the identity of the team.

In CASP, blind testing of claimed "best methods" led to a more objective discussion of the merits of existing methods, something that we expect to promote with IMPROVER.