sbv IMPROVER
SYSTEMS BIOLOGY VERIFICATION

www.sbvimprover.com

The Systems Toxicology Computational Challenge:

Marker of Exposure Response Identification

Scoring results & Lessons learned

Vincenzo Belcastro, PhD

July 11th 2016, ISMB, Orlando

- Participation & Scoring methodology
- Scoring Results & Top performing teams
  - Sub-Challenge 1 (SC1)
  - Sub-Challenge 2 (SC2)
- Post-challenge Analysis – Initial results
  - Misclassifications
  - Confidence values distribution per class
  - Gene signatures co-occurrence for smoking exposure response status (Smokers vs Non-Current Smokers)
    - Human-specific signatures
    - Species-independent signatures

Conclusions

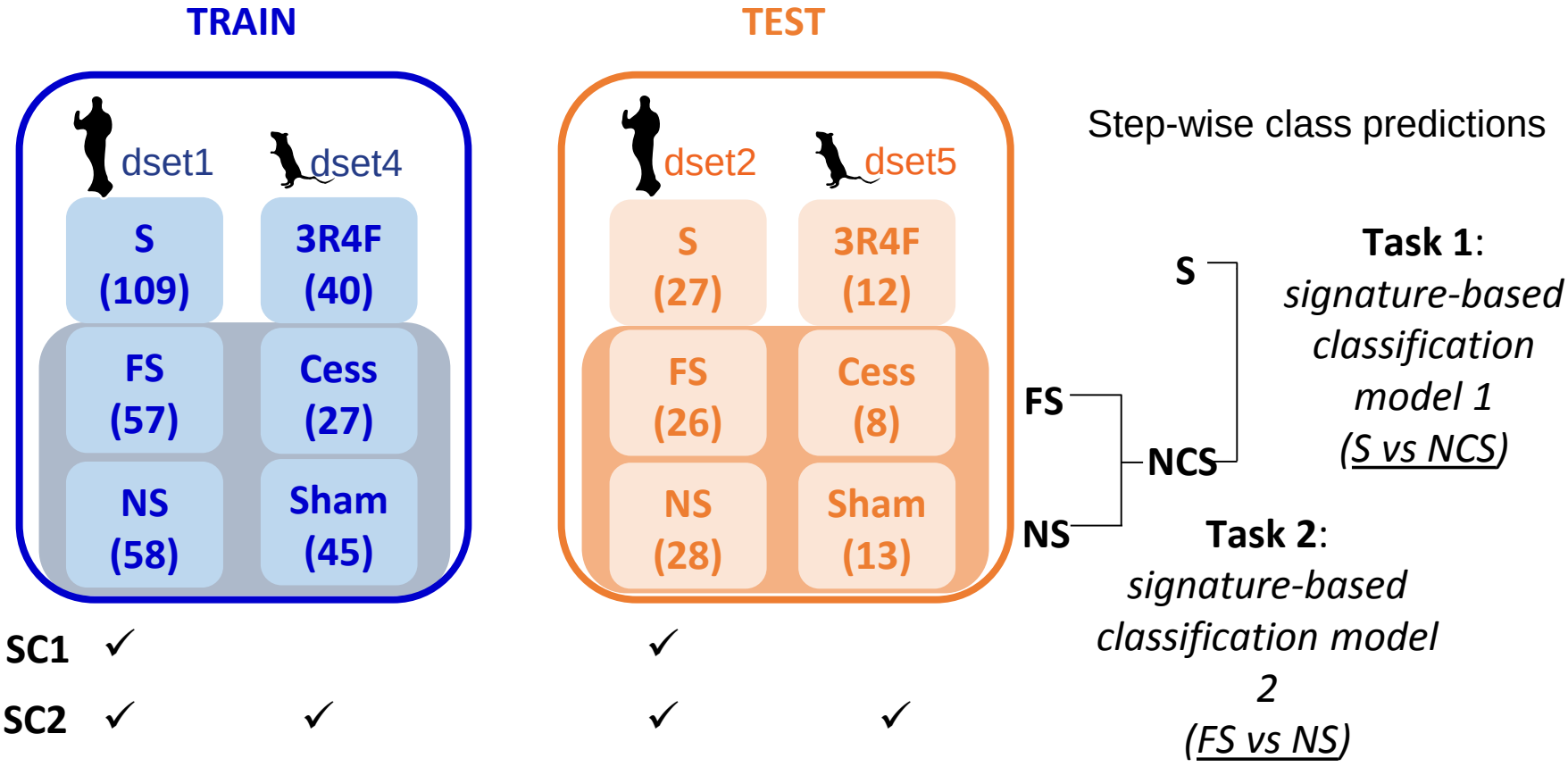
Participation & Scoring methodology

Computational Challenge Participation across the globe

135 registered participants in 61 teams – **23 teams** submitted files in at least 1 sub-challenge

Datasets provided for training and tests

Gene expression data generated from human and mouse blood samples



S/3R4F: Smokers / 3R4F (exposure to smoke from a reference cigarette)
FS/Cess: Former smokers / Cessation
NS/Sham: Never smokers / Sham
NCS: Non-current smoker

Freedom to use two separate models for 2-class prediction for each step, or directly a 3-class prediction model

https://sbvimprover.com/challenge-4/the-computational-challenge/hbs

Scoring metrics and aggregation

Task 1 predictions

Task 2 predictions

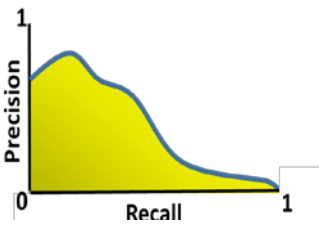
(only Task 1 True Negatives)

Confusion matrix for

Task 1		Task 2	
TP	FP	TP	FP
FN	TN	FN	TN

Area Under Precision Recall (AUPR) curve

$$precision = \frac{TP}{TP + FP}$$



$$recall\ (or\ sensitivity) = \frac{TP}{TP + FN}$$

Matthew correlation coefficient (MCC*)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

*Threshold value is 0.5

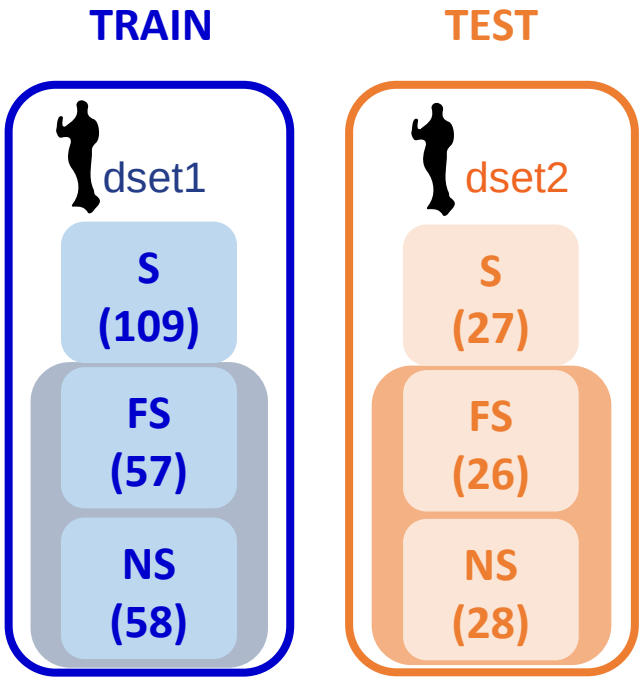
Score aggregation

- For each sub-challenge, average ranks across metrics and tasks were computed.
- Team with the lowest average rank was considered as best performer

Qualifications rules & procedure for scoring

- **Rules as per challenge description**
  - All requested files submitted:
    - 4 prediction files
    - 2 gene lists (in case of 3-class model: 1 gene list)
    - 1 write-up
  - Confidence values for common samples are the same
  - Gene list length does not exceed 40 genes

- **Scoring procedure**
  - **Anonymized** submissions
  - Results and final ranking presented to and approved by a external and independent **Scoring Review Panel**

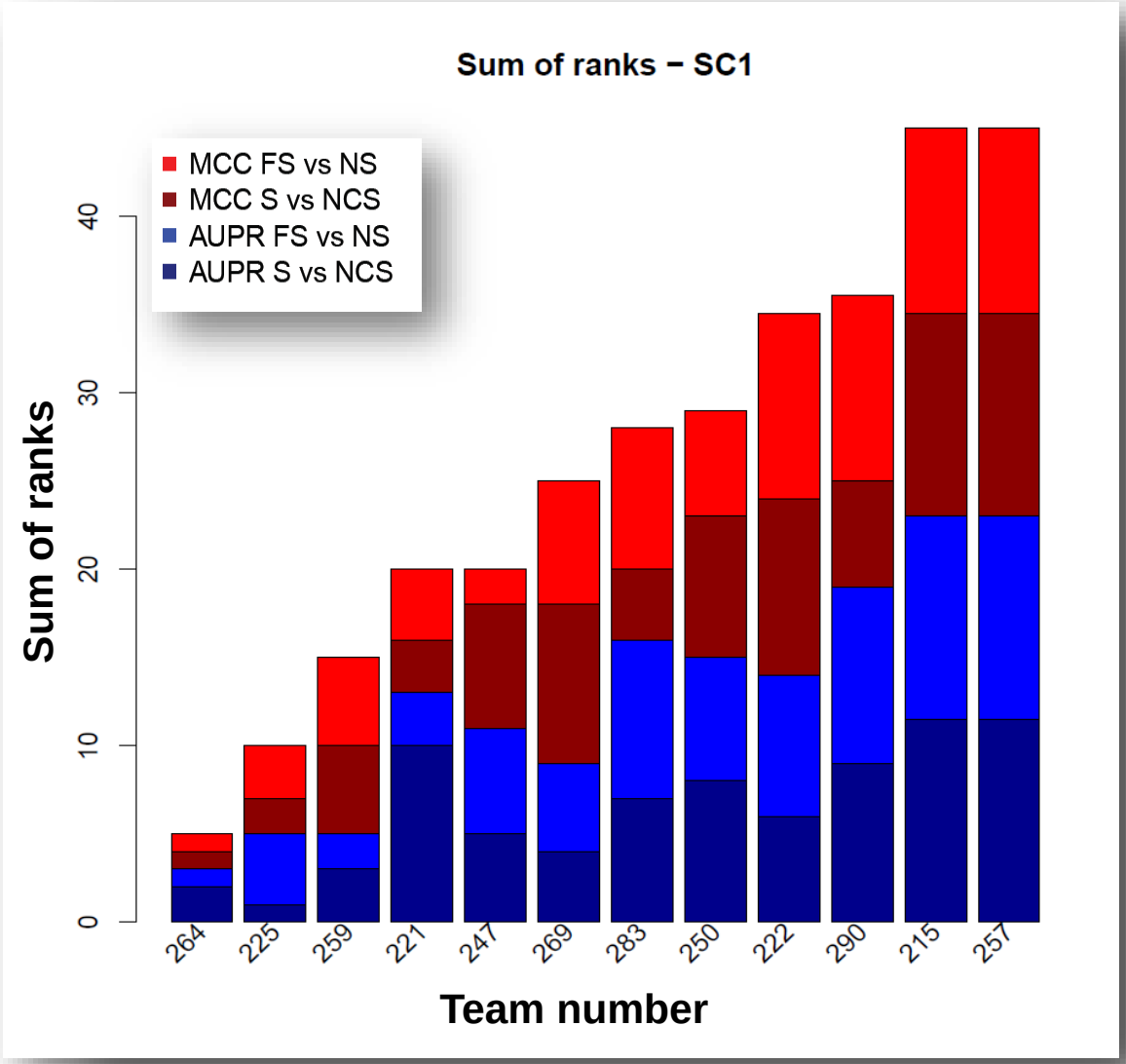


Scoring Results

Sub-challenge 1

Participants' prediction score ranking

23 teams provided submissions → **12 teams** had valid submissions



And the winners are…

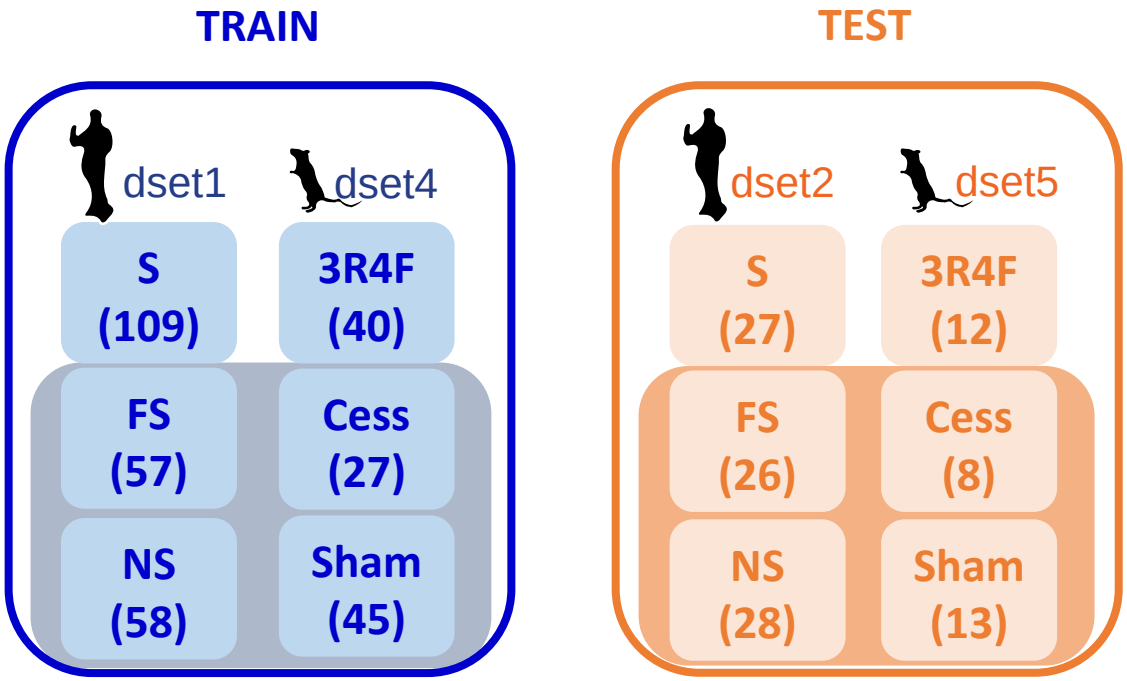
The top 3 best performing teams for Sub-Challenge 1

1. **Team 264**
2. **Team 225**
3. **Team 259**

- **Team 264**
  - Adi L Tarca — USA
  - Roberto Romero

- **Team 225**

- **Team 259**
  - Zhongqu Duan — China
  - Hao Yang
  - Xiaofeng Gong
  - Peixuan Wang
  - Chenfang Zhang
  - Wenxin Yang

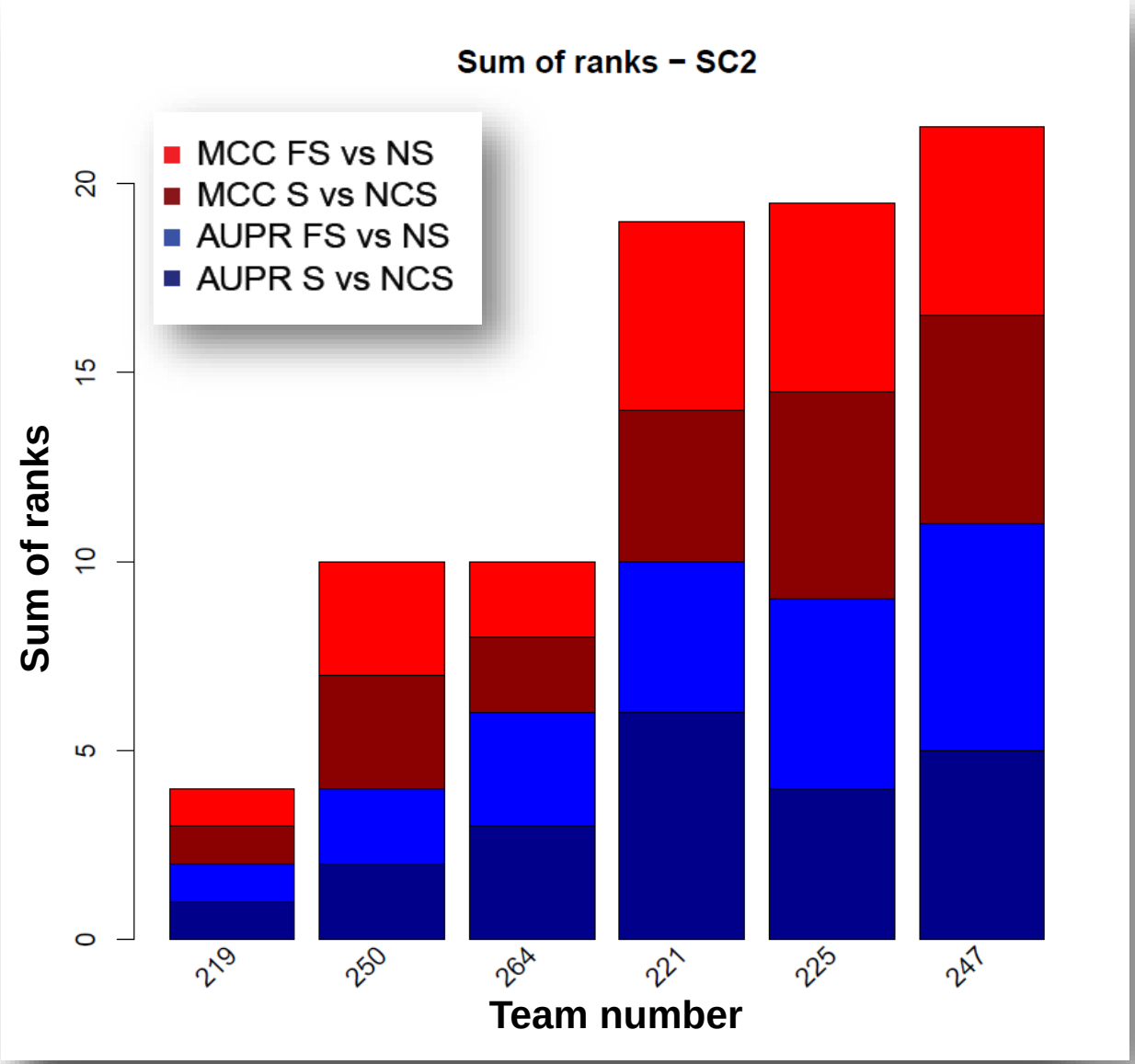


Scoring Results

Sub-challenge 2

Participants' prediction scoring

15 teams provided submissions → **6 teams** with valid submissions



And the winners are…

The top 3 best performing teams for Sub-Challenge 2

1. **Team 219**
2. **Team 250 & 264**

- **Team 219**
  - Omer Sinan Sarac — Turkey
  - Ismail Bilgen
  - Ali Tugrul Balci

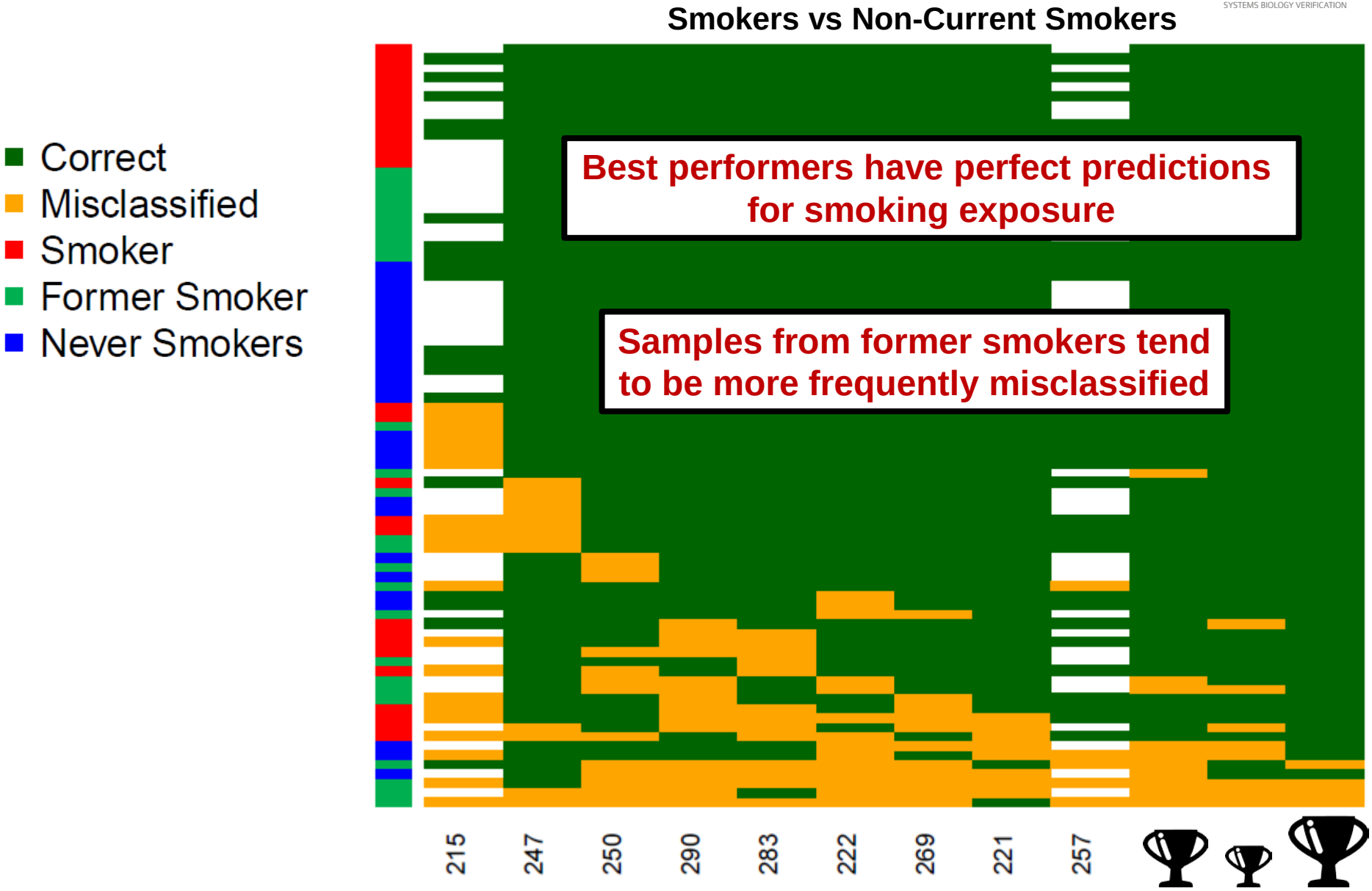
- **Team 250**
  - Rahul Kumar — USA
  - Sandeep Dhanda

- **Team 264**
  - Adi L Tarca — USA
  - Roberto Romero

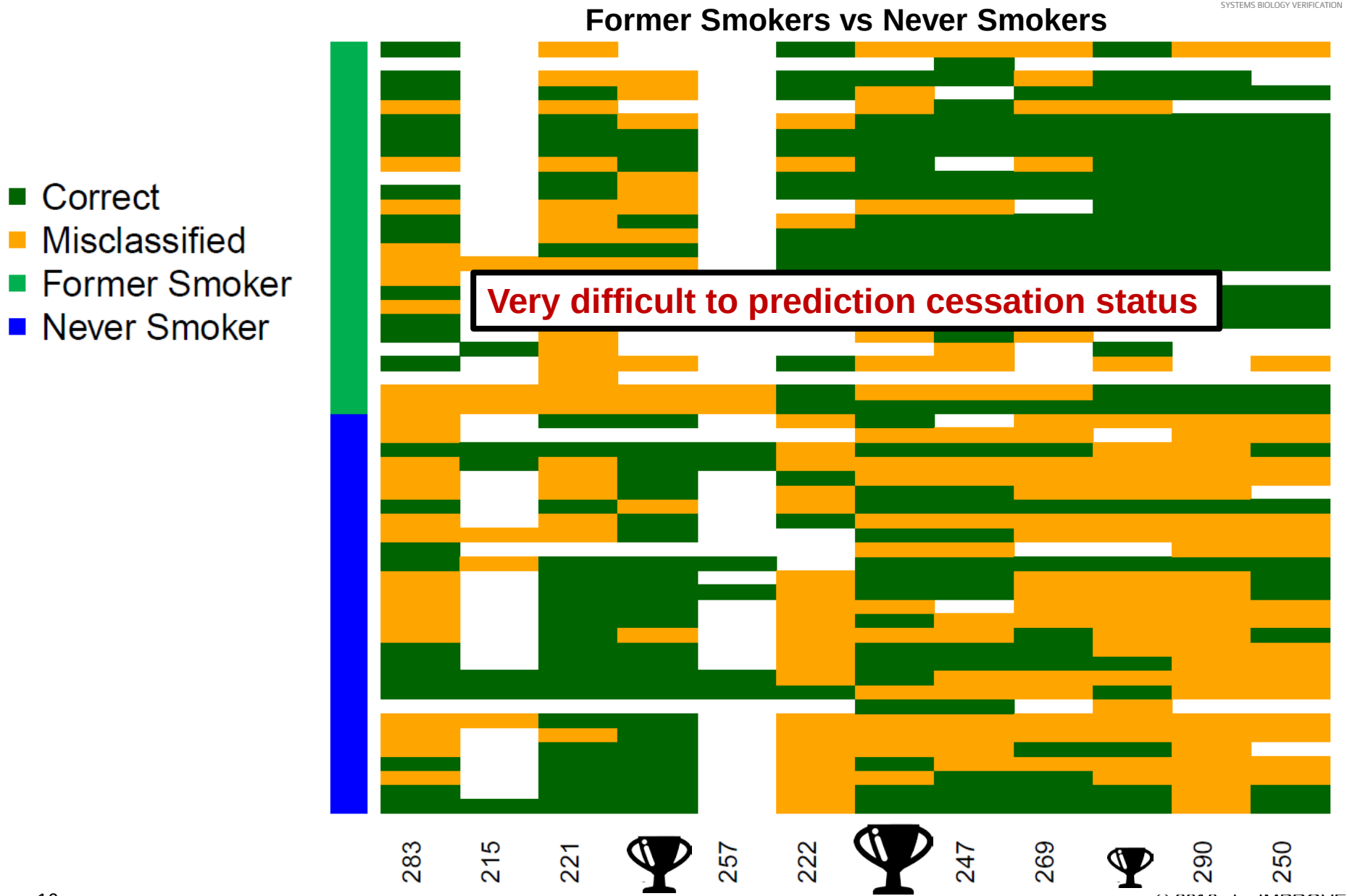
Post-challenge analysis

Misclassification sub-challenge 1

Misclassification Sub-Challenge 1



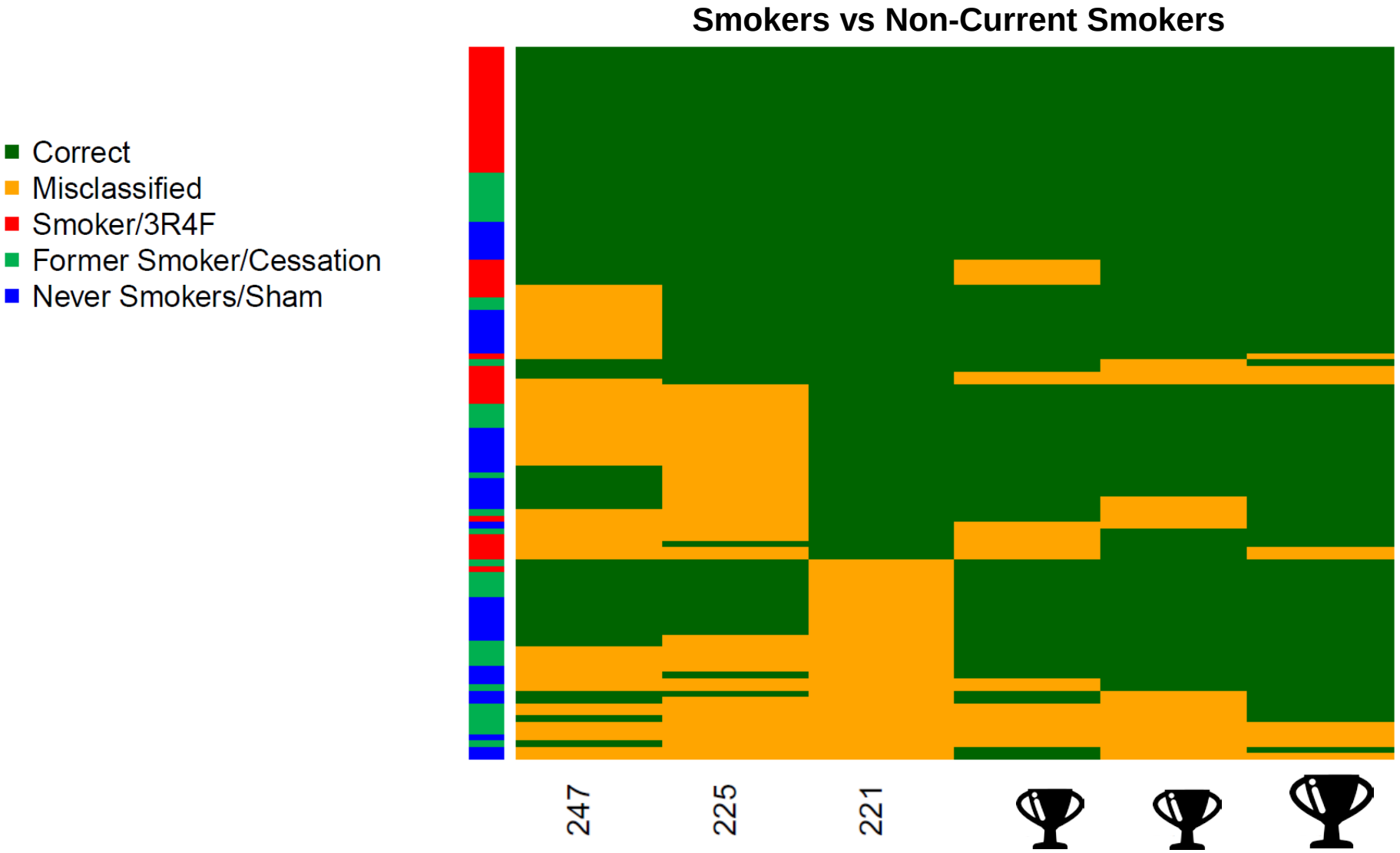
Misclassification – Sub-Challenge 1



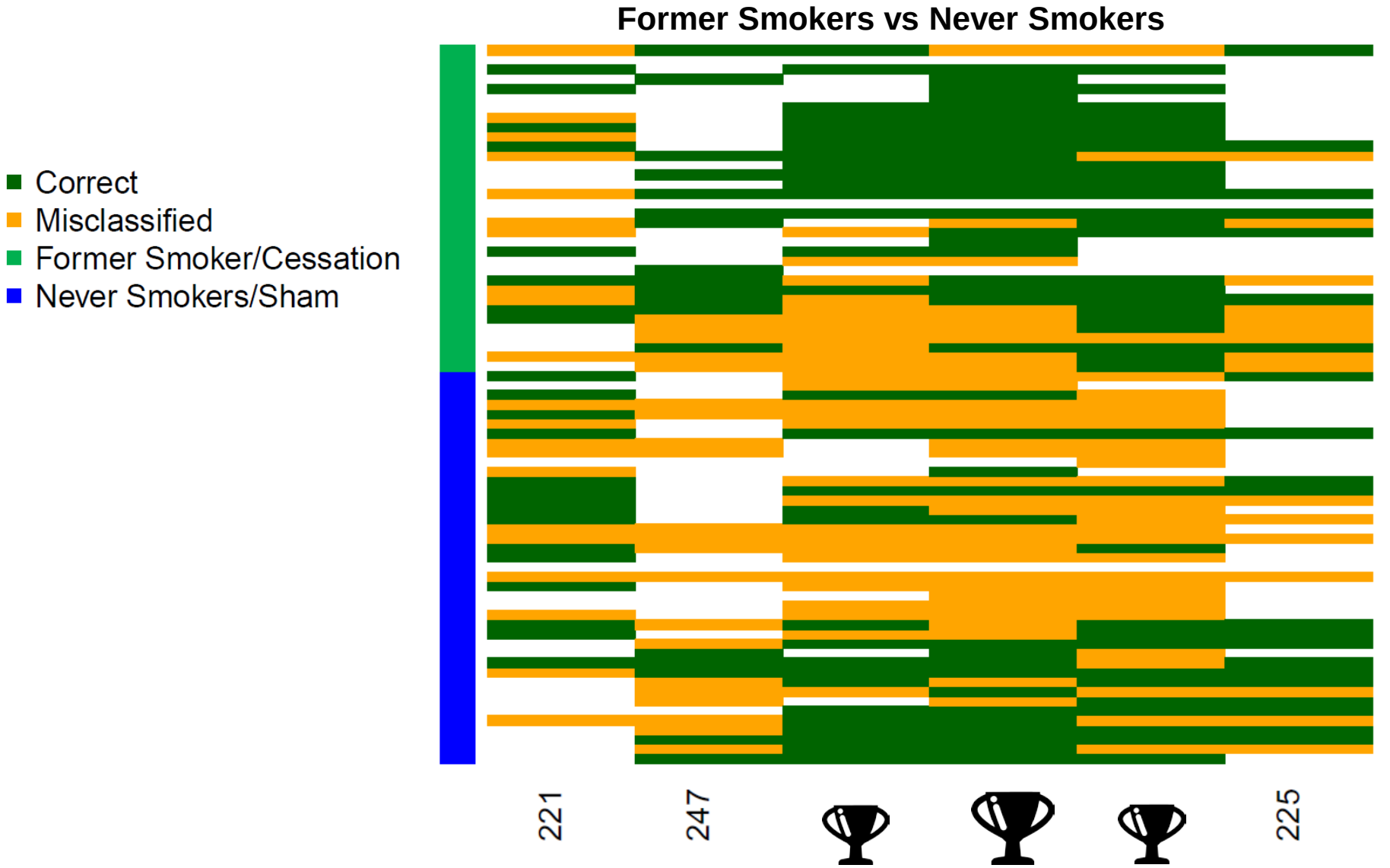
Post-challenge analysis

Misclassification sub-challenge 2

Misclassification – Sub-Challenge 2



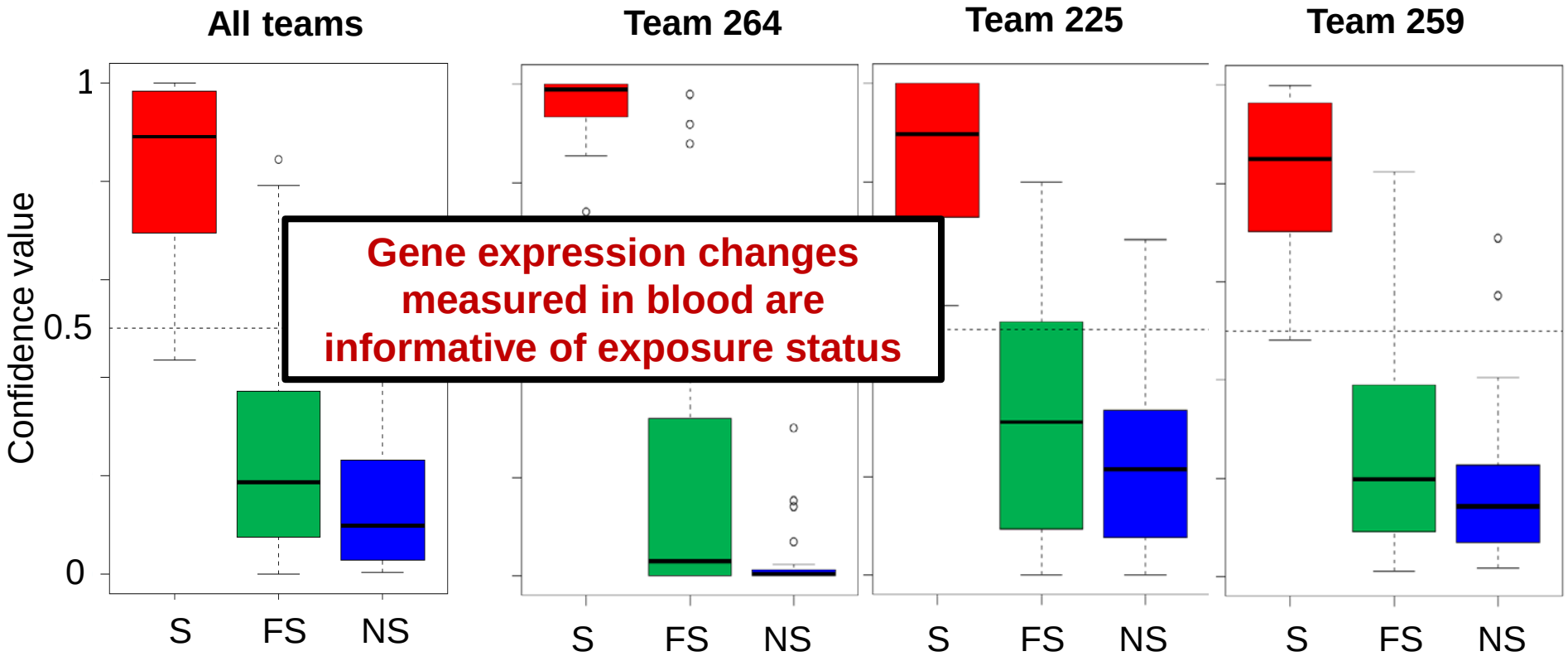
Misclassification – Sub-Challenge 2



Post-challenge analysis

Confidence values distributions per class

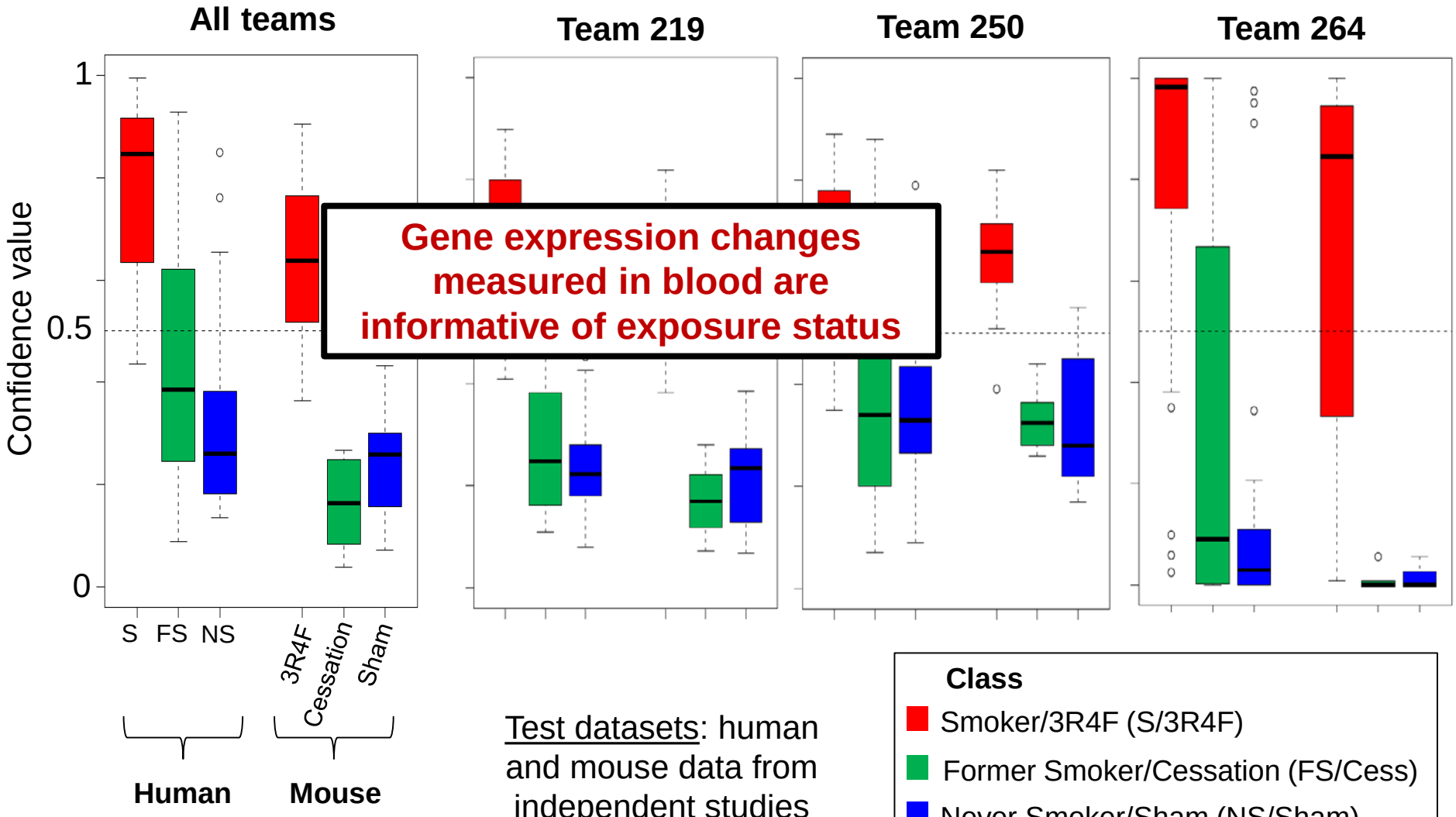
Confidence value distributions per class for all and top 3 teams
Sub-challenge 1



Test dataset: human data from an independent study

Class
- Smoker (S)
- Former Smoker (FS)
- Never Smoker (NS)

Confidence value distributions per class for all and top 3 teams
Sub-challenge 2



Post-challenge analysis

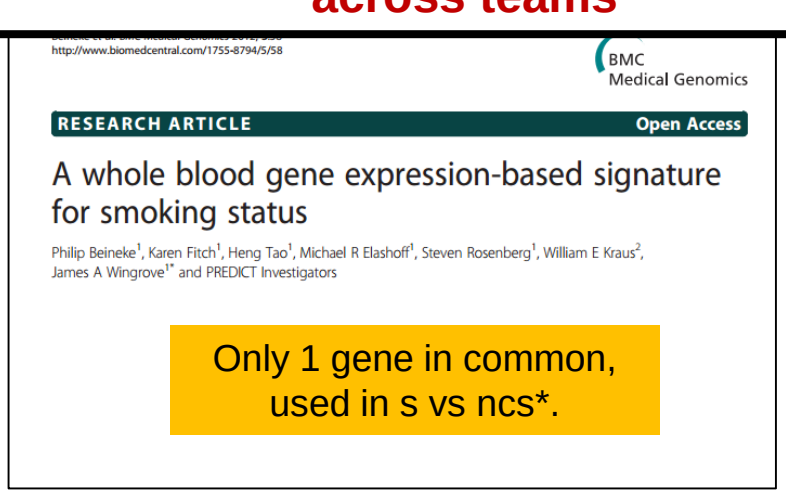
Gene signatures co-occurrence for smoking exposure response (Smokers vs Non-Current Smokers)

Human-specific gene signature for smoke exposure response Sub-Challenge 1

Team264	Team225	Team259
Gene1	Gene1	Gene1
Gene2	Gene2	Gene2
Gene3	Gene3	Gene3
Gene4	Gene4	Gene4
Gene5	Gene5	Gene5
Gene6	Gene6	Gene6
Gene7	Gene7	Gene7
Gene8	Gene8	Gene8
Gene9	Gene9	Gene13
Gene10	Gene10	Gene14
Gene11	Gene11	Gene15
Gene12	Gene	
	Gene	
	Gene	
	Gene	
	Gene	
	Gene22	
	Gene23	
	Gene24	
	Gene25	
	Gene26	
	Gene27	
	Gene28	
	Gene29	
	Gene30	
	Gene31	
	Gene32	

All teams

Human-specific smoking gene signatures include a core gene subset highly consistent across teams


http://www.biomedcentral.com/1755-8794/5/58

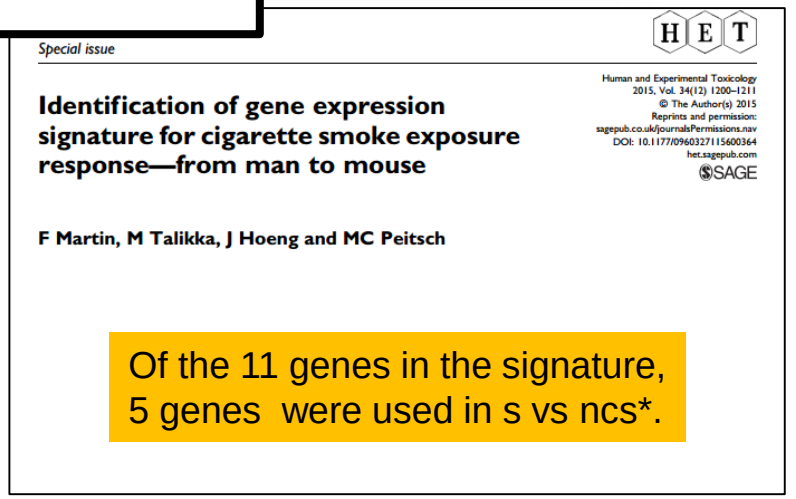
BMC Medical Genomics

RESEARCH ARTICLE Open Access

A whole blood gene expression-based signature for smoking status

Philip Beineke[1], Karen Fitch[1], Heng Tao[1], Michael R Elashoff[1], Steven Rosenberg[1], William E Kraus[2], James A Wingrove[1*] and PREDICT Investigators

Only 1 gene in common, used in s vs ncs*.


Special issue

HET

Human and Experimental Toxicology 2015, Vol. 34(12) 1200–1211 © The Author(s) 2015 Reprints and permission: sagepub.co.uk/journalsPermissions.nav DOI: 10.1177/0960327115600364 het.sagepub.com

SAGE

Identification of gene expression signature for cigarette smoke exposure response—from man to mouse

F Martin, M Talikka, J Hoeng and MC Peitsch

Of the 11 genes in the signature, 5 genes were used in s vs ncs*.

** Matching based on gene name*

Conclusions

- Successful **worldwide participation** to the challenge
- Gene expression changes measured in **blood are informative of exposure status:**
    1. Identification of **smoking exposure** status is **possible** (S vs NCS)
    2. Identification of **cessation status** is more **challenging** (FS vs NS)
- Participants succeeded in development of **inductive classification models**
- Random forest, linear discriminant analysis, partial least square discriminant analysis and logistic regression were machine learning methods used by best performing teams
- Samples from **former smokers** tend to be more **frequently misclassified**
- Exposure response markers predictive of smoking status were **identified in human blood** and included a core gene subset **highly consistent** across teams

Acknowledgements

sbv IMPROVER team

- Vincenzo Belcastro
- Filipe Bonjour
- Stephanie Boue
- Laure Cannesson
- Anouk Ertan
- Sylvain Gubian
- Julia Hoeng
- Florian Martin
- David Page
- Manuel Peitsch
- Carine Poussin
- Alain Sewer
- Marja Talikka
- Bjoern Titz
- Yang Xiang

Scoring Review Panel

- Prof. Leonidas Alexopoulos (National Technical University of Athens)
- Prof. Rudiyanto Gunawan (ETH Zurich)
- Dr. Alberto de la Fuente (Leibniz Institute for Farm Animal Biology)

To learn more, visit sbvimprover.com/comp-start

Questions? Contact Us sbvimprover.RD@pmi.com

The sbv IMPROVER project, the websites and the Symposia are part of a collaborative project designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results. The project is led and funded by Philip Morris International. The current challenges, website and biological network models were developed and are maintained as part of a collaboration with Selventa, Douglas Connect, SBX-Garuda, Nebion, OrangeBus and ADS. For more information on the focus of Philip Morris International's research, please visit www.pmiscience.com.