COSMO: A precisionFDA NCI-CPTAC Sample Mislabeling Challenge Project

Seungyeul Yoo[1,#], Zhiao Shi[2,#], Bo Wen[2,#], SoonJye Kho[3,#], Weiping Ma[1], Zeke Maier[4], Elaine Johanson[5], Henry Rodriguez[6], Jun Zhu[1], Emily Boja[5,6], Pei Wang[1,*], Bing Zhang[2,*]

# Equal contribution; * Co-corresponding (Contacts: pei.wang@mssm.edu, bing.zhang@bcm.edu)

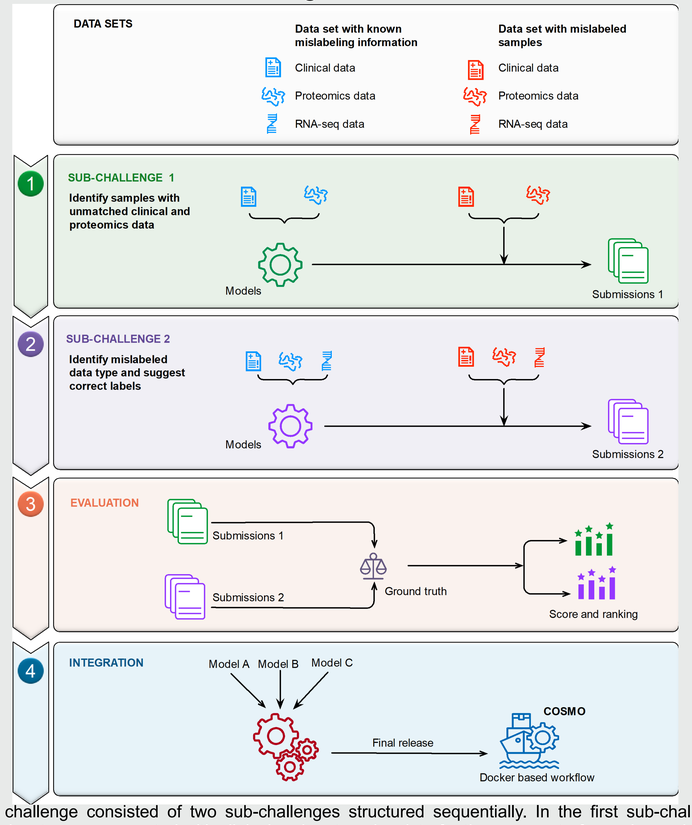
1. Icahn School of Medicine at Mount Sinai, 2. Baylor College of Medicine, 3. Wright State Univ, 4. Booz Allen, 5. FDA, 6. NCI-CPTAC

Abstract

Sample mislabeling errors are inevitable in large scale, multi-omics data during the generation, management, analysis and contributes to irreproducible results and invalid conclusions. The FDA and NCI-CPTAC launched a challenge to provide a framework for systematic benchmarking and evaluation of mislabel identification and correction methods. The challenge received a large number of submission from domestic and international scientists, with highly variable performance observed across the submitted methods. Post-challenge collaboration between the top performing teams and the challenge organizers created an open source software, COSMO, which demonstrated high accurate and robustness in mislabeling identification and correction in simulated and real multi-omics datasets.

Challenge overview

Figure 1. Overview of FDA challenge



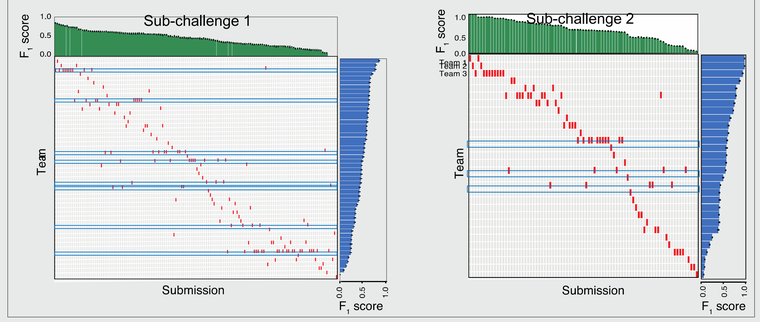
The challenge consisted of two sub-challenges structured sequentially. In the first sub-challenge, participants were presented with clinical and proteomics data for the same set of samples and asked to detect samples with unmatched clinical and proteomics data. In the second sub-challenge, participants were further provided with RNA-seq data for the same samples as in the first sub-challenge, and were requested to detect the mislabeled samples, identify the problematic data types, and to correct the errors. F_1 scores were used for performance evaluation. In the end, the top performing teams worked together to develop and implemented an automated sample labeling check algorithm named COSMO (COrrection of Sample Mislabeling by Omics).

Data preparation

Table 1. Sample labeling errors in the challenge dataset

Testing Sample	#19	#37	#6	#46	#66	#45	#13	#3	#35	#36	#38	#39	#57	#58	#59	#60	#42	#53	#9	#8
Original	s101	s98	s179	s41	s38	s153	s120	s181	s154	s18	s30	s136	s172	s142	s141	s90	s22	s7	s160	s102
Clinical	s101	s98	s179	s41	s38	s153	s120	s181	s154	s18	s30	s136	s172	s142	s141	s90	s7	s22	s102	s160
Spectra	s98	s101	s41	s179	s38	s153	s120	s181	S32 dup	s154	s18	s30	s172	s142	s141	s90	s22	s7	s160	s102
RNAseq	s101	s98	s179	s41	s153	s38	s181	s120	s154	s18	s30	s136	S13 dup	s172	s142	s141	s22	s7	s160	s102

Figure 2. Summary of challenge performance



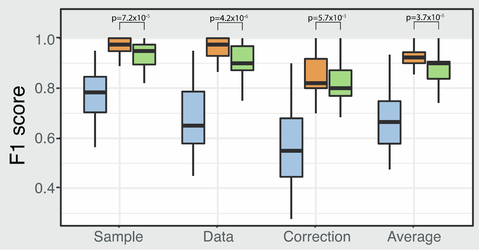
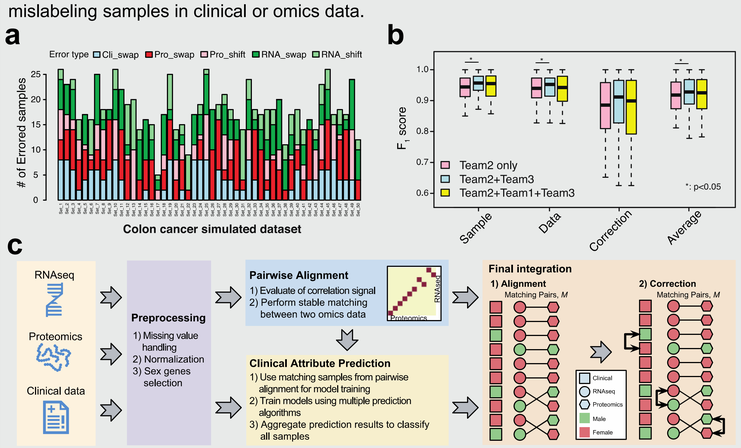


Figure 3. Evaluation of the robustness of the top three methods from sub-challenge 2 using 50 colon cancer simulated datasets with fixed types and number of errors.

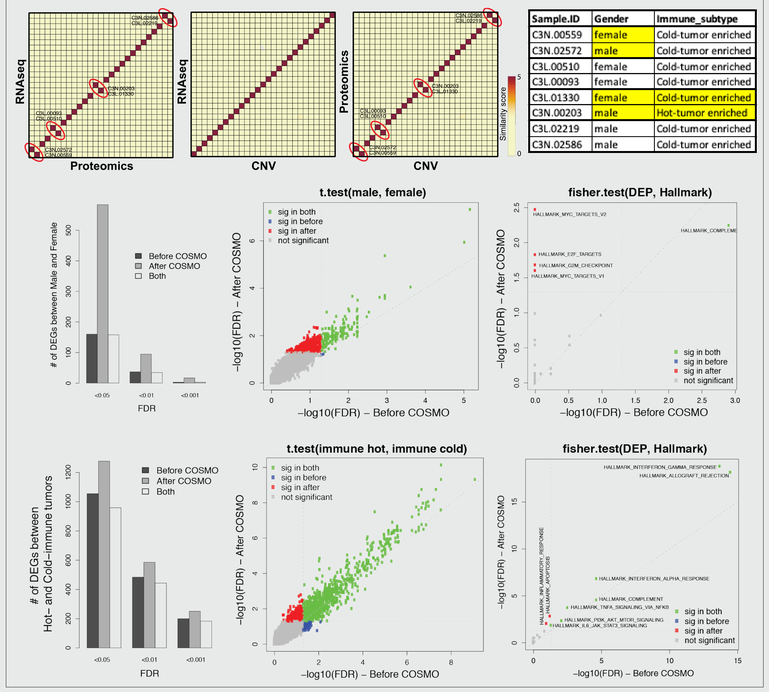
Figure 4. COrrection of Sample Mislabeling by Omics (COSMO) and its performance on independent test data sets.
a. Mimicking real cases of the sample mislabeling by generation of simulated dataset with different types and number of sample labeling errors. **b.** Performance with different sources of clinical attribute predictions. **c.** Overall schematic of the COSMO to detect and correct mislabeling samples in clinical or omics data.



Docker package of COSMO is ready-to-use and available from **https://github.com/bzhanglab/COSMO**

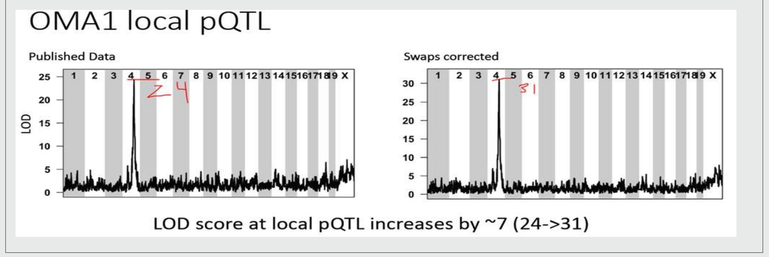
Challenge submission

Figure 5. Real dataset case 1) Four pairs of swapping in CPTAC LUAD proteomics data
- Corrected before analysis and public sharing



Sample.ID	Gender	Immune_subtype
C3N.00559	female	Cold-tumor enriched
C3N.02572	male	Cold-tumor enriched
C3L.00510	female	Cold-tumor enriched
C3L.00093	female	Cold-tumor enriched
C3L.01330	female	Cold-tumor enriched
C3N.00203	male	Hot-tumor enriched
C3L.02219	male	Cold-tumor enriched
C3N.02586	male	Cold-tumor enriched

Figure 6. Real dataset case 2) Nine pairs of swapping in mouse protein data
- Confirmed with the authors of the swapping between two multiplexes



Summary

- **Sample labeling check is not an option but a critical QC step before data analysis and public sharing.**
- **COSMO provides an automated pipeline to identify and correct labeling errors in large-scale errors.**
- **Correction of errors using COSMO has a large impact in analysis results.**
- **COSMO is implemented into a docker package and ready-to-use.**

Boja E, Tezak Z, Zhang B, Wang P, et al. (2018) Right data for right patient – a precisionFDA NCI-CPTAC Multi-omics Mislabeling Challenge. *Nature Medicine 24, 1301-1302*