Article

Patterns

A community effort to identify and correct mislabeled samples in proteogenomic studies

Highlights

- A community effort to combat sample mislabeling in multi-omic studies
- Computational solutions received show a wide range of accuracy
- The final collaborative product, COSMO, achieves high performance
- Applying COSMO to published datasets demonstrates biological impact of the tool

Authors

Seungyeul Yoo, Zhiao Shi, Bo Wen, ..., Emily Boja, Pei Wang, Bing Zhang

Correspondence

pei.wang@mssm.edu (P.W.), bing.zhang@bcm.edu (B.Z.)

In brief

A community effort to combat sample mislabeling in multi-omic studies leads to an open-source software, COSMO, with demonstrated high accuracy and robustness in mislabeling identification and correction in simulated and real multi-omic datasets.

Yoo et al., 2021, Patterns 2, 100245
May 14, 2021
https://doi.org/10.1016/j.patter.2021.100245

CellPress

Patterns

CellPress
OPEN ACCESS

Article

A community effort to identify and correct mislabeled samples in proteogenomic studies

Seungyeul Yoo,[1,2,3,15] Zhiao Shi,[4,5,15] Bo Wen,[4,5,15] SoonJye Kho,[6,15] Renke Pan,[7] Hanying Feng,[7] Hong Chen,[7] Anders Carlsson,[8,9] Patrik Edén,[8] Weiping Ma,[1,2] Michael Raymer,[6] Ezekiel J. Maier,[10] Zivana Tezak,[11] Elaine Johanson,[12] Denise Hinton,[13] Henry Rodriguez,[14] Jun Zhu,[1,2,3] Emily Boja,[14] Pei Wang,[1,2,*] and Bing Zhang[4,5,16,*]

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[2]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[3]Sema4, a Mount Sinai Venture, Stamford, CT 06902, USA
[4]Lester and Sue Smith Breast Center, Baylor College of Medicine, Houston, TX 77030, USA
[5]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA
[6]Wright State University, Dayton, OH 45453, USA
[7]Sentieon Inc., San Jose, CA 95134, USA
[8]Computational Biology & Biological Physics, Lund University, Lund 221-00, Sweden
[9]Bionamic AB, Lund 221-00, Sweden
[10]Booz Allen Hamilton, McLean, VA 22102, USA
[11]Office of In Vitro Diagnostics and Radiological Health, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD 20993, USA
[12]Office of Health Informatics, Office of the Chief Scientist, Office of the Commissioner, US Food and Drug Administration, Silver Spring, MD 20993, USA
[13]Office of the Chief Scientist, Office of the Commissioner, US Food and Drug Administration, Silver Spring, MD 20993, USA
[14]Office of Cancer Clinical Proteomics Research, Center for Strategic Scientific Initiatives, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA
[15]These authors contributed equally
[16]Lead contact
*Correspondence: pei.wang@mssm.edu (P.W.), bing.zhang@bcm.edu (B.Z.)
https://doi.org/10.1016/j.patter.2021.100245

THE BIGGER PICTURE Sample mislabeling is a well-recognized problem in scientific research, particularly prevalent in large-scale, multi-omic studies, due to the complexity of multi-omic workflows. Here, we describe a crowdsourced precisionFDA NCI-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge, which provides a framework for systematic benchmarking and evaluation of mislabel identification and correction methods for integrative proteogenomic studies. Individual solutions submitted by the challenge participants, even those from the same team, show a wide range of accuracy, underscoring the importance of the benchmarking effort. Post-challenge collaboration between the top-performing teams and the challenge organizers has created an open-source software, COSMO, with demonstrated high accuracy and robustness in mislabeling identification and correction in simulated and real multi-omic datasets.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Sample mislabeling or misannotation has been a long-standing problem in scientific research, particularly prevalent in large-scale, multi-omic studies due to the complexity of multi-omic workflows. There exists an urgent need for implementing quality controls to automatically screen for and correct sample mislabels or misannotations in multi-omic studies. Here, we describe a crowdsourced precisionFDA NCI-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge, which provides a framework for systematic benchmarking and evaluation of mislabel identification and correction methods for integrative proteogenomic studies. The challenge received a large number of submissions from domestic and international data scientists, with highly variable performance observed across the submitted methods. Post-challenge collaboration between the top-performing teams and the challenge organizers has created an open-source software, COSMO, with demonstrated high accuracy and robustness in mislabeling identification and correction in simulated and real multi-omic datasets.

INTRODUCTION

Recent advances in high-throughput omics technologies have enabled system-wide characterization of biological samples at different molecular levels.[1–3] For example, the National Cancer Institute (NCI)'s The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) have molecularly profiled large sets of tumors spanning the major human cancer types using genomic, epigenomic, transcriptomic, and proteomic platforms.[4–6] The resulting multi-omic data, together with associated clinical data, have greatly expanded our understanding of cancer biology and have led to new therapeutic insights into different cancer types.[7–9] As the volume and complexity of data continue to increase, unfortunately sample or data-labeling errors often occur during the process of data generation and management due to human errors. Although such errors have been a long-standing problem that contributes to irreproducible results and invalid conclusions,[10–12] they become particularly prevalent in large-scale omic studies. Indeed, sample-mislabeling problems have been observed in several CPTAC projects during data quality control steps, and considerable efforts have been made to correct these issues before public data release.

Several methods have been developed to screen for mislabeled samples through matching their genetic and genomic profiles.[13–18] Customized for genetic and genomic data, these methods have not been applied to or tested in proteomic profiles, which pose rather different data properties. For example, while mRNA levels of sex-chromosome genes such as *XIST* or *RPS4Y1* are unambiguous in inferring gender of samples,[13,19] predicting gender using proteomic data is more challenging mostly due to low coverage of sex-chromosome genes and higher noise of proteomic data. No previous studies provide a robust solution for gender inference based on proteomic data. As another example, correlation between copy number and mRNA expression has been used to detect sample mislabeling,[13,14] but it is unclear whether the same approach works when mapping mRNA and protein profiles due to the moderate correlation between mRNA and protein levels.[4] In addition, most of the existing methods focus only on error detection. The few offering correction of labeling errors require manual inspection, and thus cannot be easily scaled up or adopted by other research teams. To bridge these gaps, the precisionFDA and NCI-CPTAC called upon the scientific community at large to develop computational methods that detect and correct potential mislabeled samples in proteogenomic datasets through the "Multi-omics Enabled Sample Mislabeling Correction Challenge."[20] The top-performing algorithms resulting from the challenge have been systematically evaluated and collaboratively improved, leading to an integrated and automated open-source tool that can be broadly adopted to tackle the mislabeling problem in proteogenomic studies.

RESULTS

Description of the challenge

The challenge dataset was generated using RNA sequencing (RNA-seq), mass spectrometry-based proteomic data, and associated clinical information from two colorectal cancer studies containing a total of 181 colorectal tumor samples.[4,8] From the merged dataset, we first created 50 pairs of training and testing datasets with 80 samples each from random sampling of 160 samples (experimental procedures). In each training or testing dataset, four samples were randomly selected and assigned misannotated clinical information including gender and microsatellite instability (MSI) status; and RNA-seq or proteomic profiles of another eight samples were randomly selected and shuffled or mislabeled (experimental procedures). One pair of training/test datasets with the median difficulty level according to performances of our baseline method on these datasets was selected and used in the challenge (Figure S1). Participants were asked to explore the training dataset to learn about the features of the errors in order to detect and correct labeling errors from the testing dataset. The remaining training/test datasets were used later for post-challenging investigations.

The challenge consisted of two sub-challenges structured sequentially (Figure 1). In the first sub-challenge, participants were presented with clinical and proteomic data of the same set of samples and asked to detect samples with unmatched clinical and proteomic data. In the second sub-challenge, participants were further provided with RNA-seq data for the same set of samples as in the first sub-challenge. Assuming errors occurred in only one data type, participants were further requested not only to detect the problematic samples (level 1) but also to identify the mislabeled data types (level 2) and correct the errors (level 3). F_1 scores, i.e., harmonic means of the precision and recall of the models, were used for performance evaluation in both sub-challenges. Especially in the second sub-challenge, F_1 scores from the three levels were averaged for performance evaluation (experimental procedures).

Challenge results

A total of 52 teams from 15 countries participated in the challenge (Figure 2A), with 149 solutions submitted for sub-challenge 1 and 87 solutions for sub-challenge 2. The large number of submissions for both sub-challenges suggests the practical significance of and, thus, great interest in solving the problems in the scientific community. A striking observation from the challenge performances is that individual solutions showed a wide range of accuracy in both sub-challenges (Figures 2B and 2C for sub-challenge 1 and 2, respectively; Table S1). In some cases, even multiple solutions submitted by the same team had a wide range of accuracy (Figures 2B and 2C for sub-challenge 1 and 2, respectively). These results highlight the importance of systematic benchmarking efforts and the need for a standardized, accurate, and open-source method for mislabeling check.

A unique challenge in working with proteomic data is the presence of a significant amount of missing values. Most participants performed an imputation step to deal with this issue (Figure 2D). One frequently used approach among submitted solutions was to discard features containing missing values or extremely low values, and the teams using this approach tended to have a relatively poorer performance in sub-challenge 1 compared with the other teams (average percentile rank [APR] = 0.423; APR has a range from 0 to 1—larger is better, 0.5 is neutral). Another frequently used strategy was to replace missing values with 0.

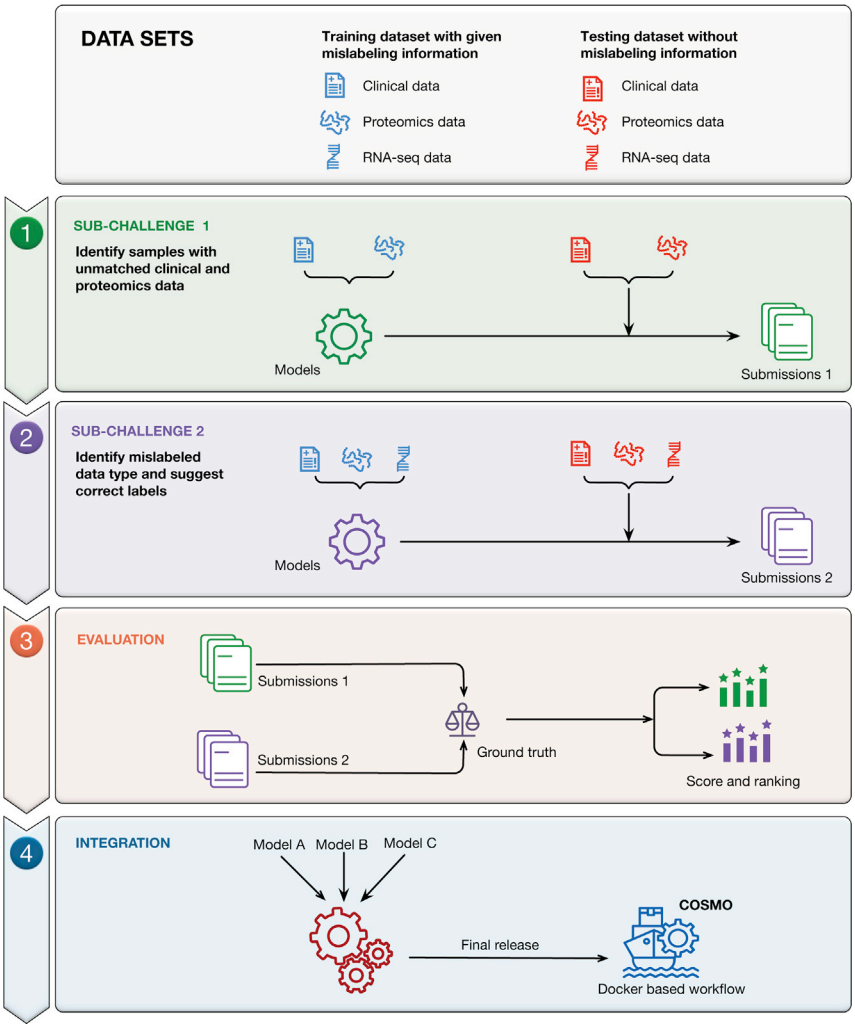

Figure 1. Overview of pFDA-NCI-CPTAC Challenge design and post-challenge development

The challenge consisted of two sub-challenges structured sequentially. In the first sub-challenge, participants were presented with clinical and proteomic data for the same set of samples and asked to detect samples with unmatched clinical and proteomic 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 correct the errors. F_1 scores were used for performance evaluation. In the end, the top-performing teams worked together to develop and implement an automated sample-labeling check algorithm named COSMO (COrrection of Sample Mislabeling by Omics).

The teams using this approach tended to have better performance compared with the other teams (APR = 0.658), and this approach was used by one of the top-performing teams to achieve a team average F_1 score of 0.83 (Figure 2D). Several teams replaced missing values with gene-wise mean or median, which assumes that the expression levels of a gene/protein in different experiments are constant. This approach tended to underperform (APR = 0.231). Model-based imputation methods, such as *k*-nearest neighbors (KNN), random forest (RF), and non-negative matrix factorization (NMF), have also been used with varied levels of success. One of the top-performing teams used NMF and achieved an F_1 score of 0.75, but generalizability of this approach is uncertain because it was used by only one team.

For sub-challenge 1 the F_1 scores ranged from 0 to 0.8, suggesting difficulty in predicting gender or MSI status based only on proteomic data (Figure 2B). Matching clinical annotations (gender and MSI) with omic data often involves constructing prediction models for clinical variables based on omic data. A summary of prediction models employed by all teams is presented in Table S2. Modeling methods used by the top-performing teams included logistic regression (LR), RF, and KNN. One important aspect of prediction model construction based on high-dimensional data is feature selection. The participating teams used a variety of feature-selection techniques (Figure 2E). The simplest approach was to remove features with variance below some threshold. The second approach picks features based on results of univariate statistical tests, such as the traditional ANOVA test or differential test developed for gene expression data analysis. Model-based feature selection, such as regularized LR (with L1 penalty), RF, and nearest shrunken centroids, were popular choices. Interestingly, the same feature-selection approach may lead to very different performances, which might be explained by their combination with different modeling methods. Many teams leveraged domain knowledge to guide the selection of important features, such as using genes from sex chromosomes to predict gender information, and the results were mixed (APR = 0.534).

For sub-challenge 2, the average F_1 scores also ranged widely from 0.1 to 0.99 (Figure 2C). For matching protein and RNA-seq data, either Pearson- or Spearman-based correlation analysis was utilized by most teams, including all three top-performing teams. A few teams preceded the correlation analysis with regression analysis that used one data modality to predict the other, which did not yield superior performance. For the final label correction, teams typically searched for patterns consistent with mismatching scenarios in different data modalities through heatmap visualization. It is worth noting that the top-performing solutions in sub-challenge 2 were able to identify mislabeled samples with much higher accuracies than those in sub-challenge 1. This clearly demonstrates the benefit of using multi-omic data for identifying sample-labeling errors.

Post-challenge collaboration and COSMO

The three top-performing teams from sub-challenge 2 were invited to participate in post-challenge collaborative

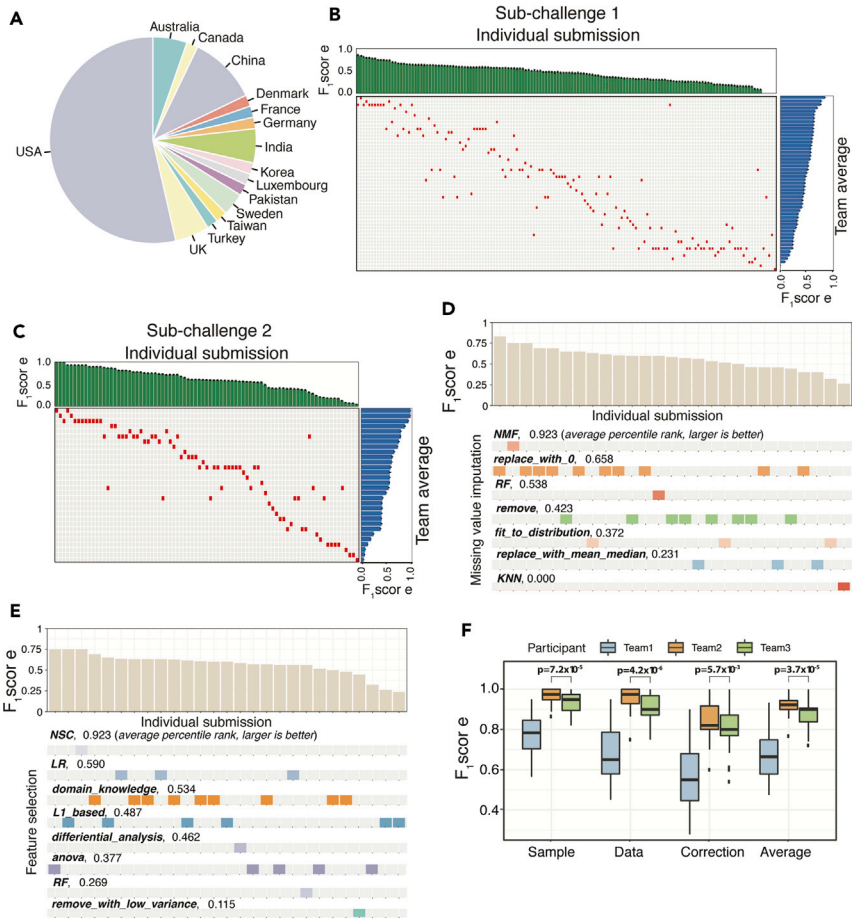

Figure 2. Summary of challenge results

(A) Global participants for the challenge suggesting high interest in the challenge problems.

(B) Performance evaluation of 149 submissions (columns) from 52 unique submitters (rows) for sub-challenge 1. The F_1 score with 95% confidence interval was evaluated for each submission and averaged for unique submitters.

(C) Evaluation of sub-challenge 2. In total 57 submissions (columns) from 31 unique submitters (rows) were evaluated in terms of average F_1 score. Wide distribution of performance of submission for both sub-challenges was observed. Even within the same team, performance varied in a wide range, suggesting significance of standardized methods.

(D and E) Association between team performances in sub-challenge 1 and missing data imputation methods (C) and feature-selection methods (D). Metric used: average percentile rank.

(F) 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. P values were calculated using two-sided paired Student's t test.

development. First, we further evaluated the robustness of the three winning methods (supplemental experimental procedures) by applying them to the original 50 training/testing datasets from which the challenge dataset was selected (experimental procedures). Methods from both Teams 2 and 3 showed high accuracy with average F_1 scores around 0.9, and the method from Team 3 showed the best performance at all levels of evaluation (two-sided paired Student's t test, $p < 0.01$, Figure 2F). In contrast, the average F_1 score of our baseline method was only 0.68 (experimental procedures), about 30% lower than the scores of winning methods from Teams 2 and 3. The performance of the method from Team 1, however, was relatively low in general, having an average F_1 score of 0.66. This is mainly due to the difficulty for Team 1 to implement their manual inspection procedures, which was used during the challenging phase, in an automatic pipeline (Figure 2F). These results underscore the power of crowdsourcing in achieving optimal performance in mislabeling correction and suggest pipeline automation as a key factor for robust performance.

In both the challenge and the above robustness evaluation exercise, training datasets have the same patterns and frequencies of errors as the test datasets. However, in a real-world scenario, training data are not available and there is no prior information on the patterns and frequencies of mislabeling errors. To better mimic real-world applications, we generated 50 new datasets with varying mislabeling error rates and patterns based on the colon data, and the new datasets did not include matched training data (experimental procedures and Figure 3A). The pipeline from Team 2 was successfully adapted to fully address this new challenge and achieved a high median average F_1 score of 0.92 (Figure 3B). However, the pipelines from the other two teams could not be effectively adapted to detect mislabeling errors when error rates and error patterns were unknown. We further tested whether integrating intermediate clinical attribute prediction results from multiple teams, i.e., "wisdom of the crowds," could lead to better performance than the best single approach. By integrating results from Team 3 with the ones from Team 2, we observed small but significant improvement of the average F_1 scores for detecting problematic samples ($p = 0.03$), identifying mislabeled data types ($p = 0.03$), and the overall performance ($p = 0.01$) (Figure 3B). There was also a trend of increasing performance for error correction, albeit not significant ($p = 0.07$). Further integrating the results from Team 1 did not lead to additional improvement.

Based on these results, we developed an automated sample-mislabeling check pipeline named COSMO (COrrection of Sample Mislabeling by Omics) following Team 2's overall approach, but also integrated the clinical attribute prediction algorithm from Team 3 (Figure 3C and experimental procedures). For independent validation, we applied COSMO to 50 simulated datasets from a kidney cancer study (experimental procedures and Figure 3D) with varying error rates and patterns, and obtained a median average F_1 score of 0.99 (Figure 3E), demonstrating high accuracy and robustness of the COSMO pipeline. We also associated the error rate with the performance of COSMO in both colon and kidney datasets (Figure S2). As expected, COSMO showed better performance with lower error rates while

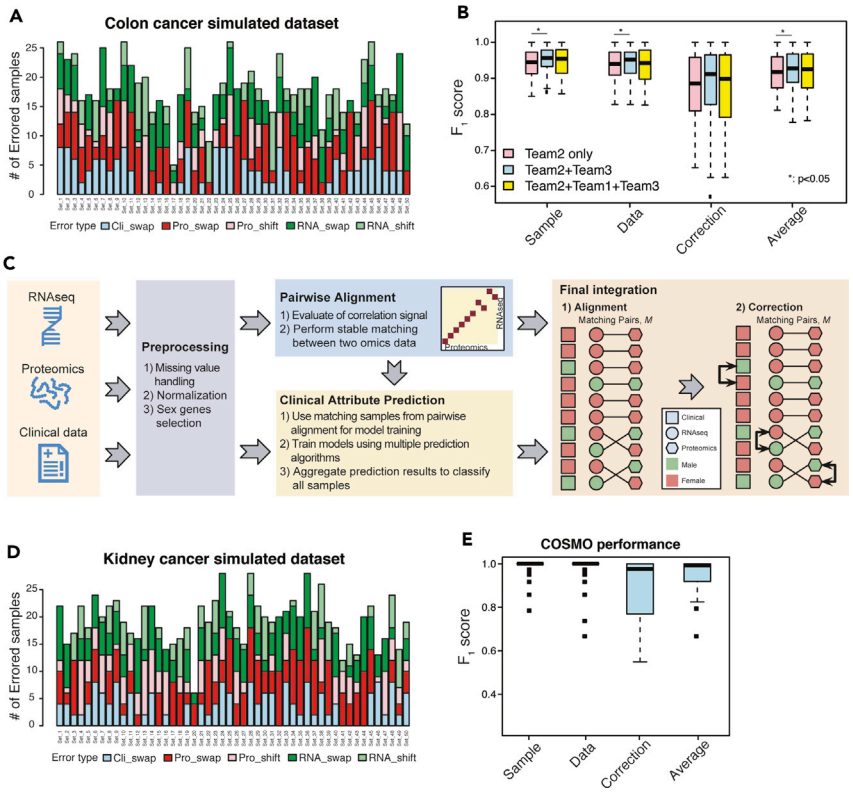

Figure 3. COSMO and its performance on independent test datasets

(A) Mimicking real cases of the sample mislabeling by generation of simulated dataset with different types and number of sample-labeling errors from the colon cancer dataset.

(B) Performance with different sources of clinical attribute predictions. P values were calculated using two-sided paired Student's t test.

(C) Overall schematic of COSMO to detect and correct mislabeling samples in clinical or omic data.

(D) Mimicking real cases of sample mislabeling by generation of simulated dataset with different types and number of sample-labeling errors using CPTAC kidney cancer datasets.

(E) Performance of COSMO in the 50 simulated datasets from (D).

high accuracies (F_1 score > 0.9) were still achieved with relatively high error rates (>20%). In the kidney cancer dataset, COSMO's performance was almost perfect for the cases with error rate below 20%.

Application of COSMO to real-case datasets

To test COSMO's performance in real multi-omic studies, we applied it to six independent multi-omic datasets (experimental procedures and Table S3). First, we applied COSMO to three human tumor datasets in which mislabeled samples were observed previously either before or after publication: the pre-quality control (preQC) CPTAC lung cancer dataset (preQC CPTAC LUAD),[21] the preQC CPTAC kidney cancer dataset (preQC CPTAC CCRCC),[7] and the TCGA breast cancer dataset (TCGA BRCA)[6] (experimental procedures). Applying COSMO to the preQC CPTAC LUAD dataset identified four pairs of swapping samples in the proteomic data by integrating results from RNA-seq-Proteomics, RNA-seq-CNV (copy-number variation), and Proteomics-CNV alignments (Figure 4A). In the preQC CPTAC CCRCC dataset, the heatmaps generated by COSMO clearly revealed reciprocal mislabeling among three samples in the proteomic data (Figure 4B). In both cases, these errors were previously identified by the CPTAC data analysis centers during data quality control, confirmed by data generation centers, and consequently corrected before the final data release and publication. In the TCGA BRCA dataset, a previous study reported eight sample swaps in the microarray data.[13] COSMO recapitulated the exact same eight pairs swapped in microarray data by integrating microarray, RNA-seq, and CNV data (Figure 4C).

Next, we applied COSMO to three other published multi-omic studies for which sample mislabeling has not been reported previously. First, we investigated Cancer Cell Line Encyclopedia (CCLE)[24] data of 371 cell lines for which RNA-seq, proteomic, and CNV data are available. COSMO showed that all samples were perfectly aligned across RNA-seq, proteomic, and CNV profiles in this dataset. Next, using RNA-seq, proteomic, and Riboseq profiles of 62 human lymphoblastoid cell lines generated in a study characterizing the impact of genomic variation on RNA and protein,[22] COSMO identified a swap of two samples in RNA-seq and a potential duplicated sample in proteomic data (Figure 4D). In another study investigating how genetic variation affects transcript and protein abundance in livers from 192 Diversity outbred mice,[23] nine swapping pairs were detected by COSMO (Figure 4E). In addition, by comparing predicted sexes from RNA-seq and proteomic data with corresponding clinical annotations, COSMO attributed the labeling errors to proteomic data for four swapping sample pairs with different sexes (Figure 4E and Table S4). Further investigation of the proteomic experimental design of the study[23] revealed that the sample-labeling swapping occurred between two multiplexed tandem mass tag experiments. Following these findings reported by COSMO, the authors of the publication confirmed the sample-labeling errors in the proteomic dataset, and a request for correction has been submitted to the journal (S. Munger, personal communication). Detailed results from the six independent datasets are summarized in supplemental experimental procedures.

In summary, these results demonstrate general applicability of COSMO to sample-labeling correction in multi-omic studies involving different types of omic platforms, different organisms, and both cancer and non-cancer studies (Table S3).

Biological impact of mislabeling correction

Sample mislabeling may associate omic profiles with incorrect clinical phenotype annotations and impair differential expression analysis. Among the four swapping pairs identified in the preQC

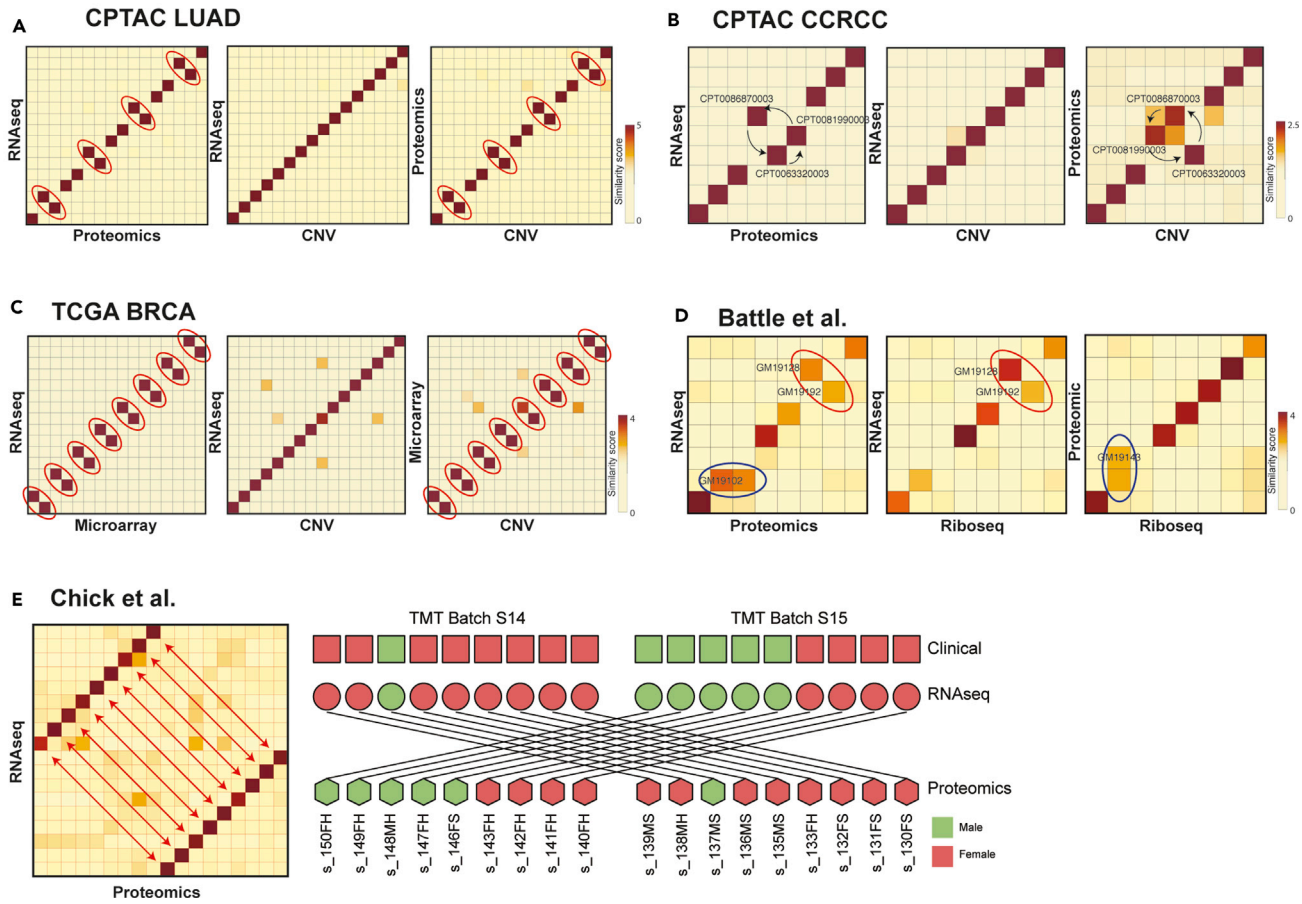

Figure 4. Application of COSMO in real datasets

(A) CPTAC LUAD: four pairs of proteomic samples reciprocally matched each other between RNA-seq-Proteomics and Proteomics-CNV, but no labeling swapping was observed in RNA-seq-CNV.

(B) CPTAC CCRCC: three samples in proteomics were shifted in RNA-seq-Proteomics and Proteomics-CNV matching while samples between RNA-seq and CNV were matched well.

(C) TCGA BRCA: eight pairs of microarray samples were swapped in RNA-seq-Microarray and Microarray-CNV matching.

(D) Battle et al.:[22] two RNA-seq samples were swapped based on alignment among RNA-seq, proteomic, and Riboseq data. Potential duplicated protein sample was observed.

(E) Chick et al.:[23] nine pairs of samples were swapped between RNA-seq and protein data. Merging with clinical annotation of gender of the sample suggested swapping in proteomic data.

proteomic data of the CPTAC LUAD study (Figure 4A), two swaps involved samples with different genders (Table S5). In the comparison between male and female samples to identify differentially expressed proteins (DEPs), based on the COSMO-corrected data, 584 DEPs were identified (Student's t test, false discovery rate [FDR] < 5%), whereas only 160 DEPS were obtained based on the preQC data (Figure 5A). The drastic difference was driven by small but meaningful changes in which mislabeling correction pushed hundreds of genes below the significance threshold (Figure 5B). The COSMO-corrected data also showed higher power in detecting gender-associated pathways, and several cell-cycle-related pathways including G2M_CHECKPINT, E2F_TARGETs, MYC_TARGETS_V1, and MYC_TARGETS_V2 could not be identified at the same significance threshold with the preQC data (Figure 5C).

Another swap in the preQC CPTAC LUAD proteomic data involved one immune-hot tumor and one immune-cold tumor. The correction of this swap is critical for the two affected patients, because it may avoid incorrect immunotherapy decisions for these patients.[21] In addition, correction of this single swapping pair had significant impact on identifying DEPs between immune-hot and immune-cold tumors. Among the 8,528 proteins in the dataset, 1,277 DEPs were identified based on the COSMO-corrected data (Student's t test, FDR < 5%), which is 20% more than the DEPs identified in the preQC data (Figure 5D). Of the 1,277 DEPs, 959 were identified in both datasets whereas 318 were identified only after mislabeling correction (Figure 5E). The COSMO-corrected data also showed higher power in detecting differential pathways (Figure 5F). Specifically, APOPTOSIS and INFLAMMATORY_RESPONSE were significantly associated with immune-hot tumors only based on COSMO-corrected data, and stronger associations were observed for other immune response-related pathways such as INTERFERON_GAMMA_RESPONSE and

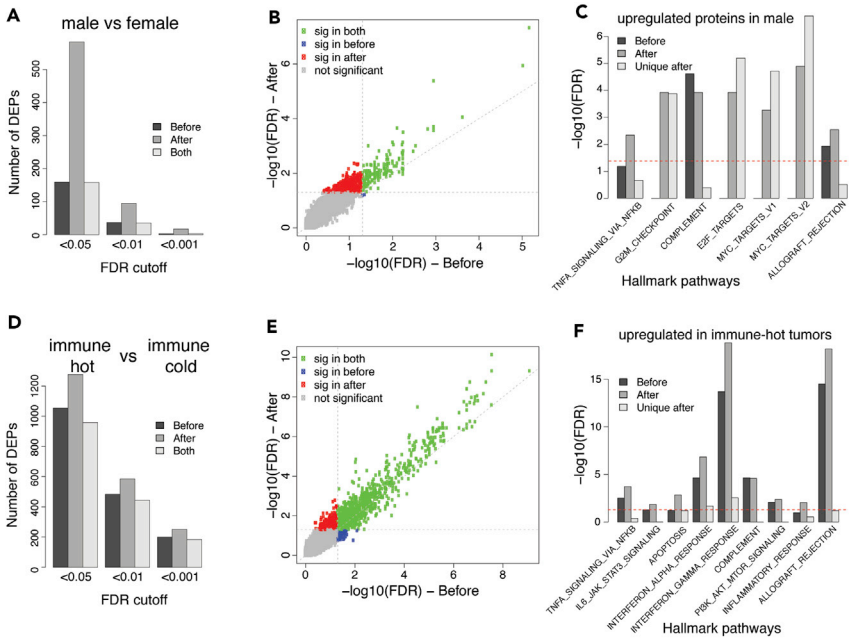

Figure 5. Biological impact of error corrections using COSMO in CPTAC LUAD dataset

(A) Number of DEPs between male and female tumors before and after error correction.

(B) Comparison of t test FDRs ($-\log_{10}$) of 8,528 proteins between male and female tumors.

(C) HALLMARK pathways (FET FDR < 0.05) significantly associated with gender DEPs before and after COSMO. Unique DEPs after COSMO were also used for functional enrichment test.

(D) Number of DEPs in immune-hot and immune-cold tumors.

(E) Comparison of t test FDRs of 8,528 proteins between immune-hot and immune-cold tumors.

(F) HALLMARK pathways significantly associated with upregulated proteins in immune-hot sub-type tumors.

(G) Correlation strengths of 8,366 gene-protein pairs before and after correction. Pearson correlation p values were adjusted as Benjamini-Hochberg adjusted p values (FDR) and then $\log_{10}$ transformed.

(H) Difference in correlation strengths of 269 gene-protein pairs significant only after error correction.

ALLOGRAFT_REJECTION.[21] These results suggest that even a small number of sample-labeling errors could have a significant impact on differential analyses at both gene and pathway levels.

Another important application of multi-omics is to investigate the relationships between different omic modalities, such as mRNA-protein correlation, expression quantitative trait loci (eQTL) analysis, and protein quantitative trait loci (pQTL) analysis. To examine the impact of sample mislabeling on assessing mRNA-protein correlation, we compared the gene-wise mRNA-protein correlations in the CPTAC LUAD study, both before and after mislabeling correction. After fixing errors in 7.5% (8/107) of the samples, COSMO-corrected data led to improved mRNA-protein correlations for about 85% of genes (Figure 5G), and 267 more genes were found to show significant RNA-protein correlation (FDR < 1%) specifically in COSMO-corrected data (Figure 5H). Several of these genes were known cancer genes, such as TBC1D15, the one with the largest change of correlation coefficient, was reported as an oncoprotein to promote self-renewal and pluripotency.[25] In addition, mRNA expression of *RAP1B* was associated with poor prognosis and promotion of an aggressive phenotype in gastric cancer.[26]

In the preQC CPTAC CCRCC dataset, the detected error rate was much lower, 3.5% (3/77). Nevertheless, we observed similar patterns of increasing mRNA-protein correlations for 62% of the genes (Figure S3A), and 54 genes only showed significant mRNA-protein correlation based on COSMO-corrected data (Figure S3B). Because discordant mRNA and protein expression is typically considered as a result of translational and protein degradation regulations, sample mislabeling may lead to overestimation of these regulations both at a global level and for individual genes.

Similarly, sample mislabeling may lead to underestimation of eQTL and pQTL effects. For the aforementioned study investigating how genetic variation affects transcript and protein abundance in livers from 192 Diversity outbred mice,[23] the authors repeated the pQTL analysis based on COSMO-corrected data and found a stronger overall impact of genetic variants on the proteome. The new analysis identified 497 more local pQTLs than in the published dataset at the same significance thresholds, and among the 1,681 local pQTLs identified in both datasets, 1,456 (87%) mapped with higher log odds ratio (LOD) scores in the updated dataset. For example, the LOD score of OMA1 local pQTL, one of the main findings in the paper,[23] increased from 24 to 31 after correction of the errors (Figure 6, S. Munger, personal communication).

Taken together, sample-labeling errors could have a significant impact on biological conclusions in omic studies, and COSMO provides an automated solution to catch and fix these errors proactively.

DISCUSSION

While integration of multiple layers of omic data is critical to provide a comprehensive understanding of molecular mechanisms underlying complex biological systems, sample mislabeling is especially prevalent in multi-omic studies and contributes to irreproducible results and invalid conclusions. Notably, although genome-wide proteomic profiling has emerged as a powerful technology in

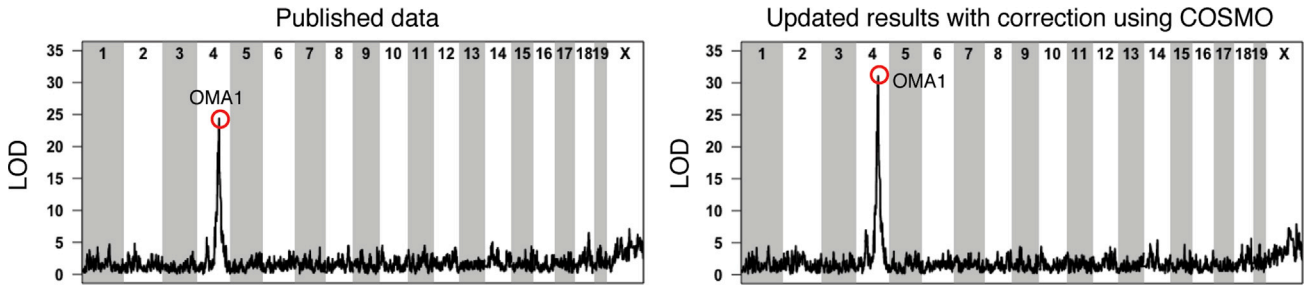

Figure 6. pQTL analysis impacted by error correction

The log odds ratio (LOD) score of OMA1 local pQTL increased from 24 (left) to 31 (right) after correction of the errors.

multi-omic studies, it remains challenging to achieve the level of sensitivity and accuracy as in RNA profiling, making it more difficult to investigate sample mislabeling in proteomic data. This study has three major contributions. First, the crowdsourcing challenge provided a framework for systematic benchmarking and evaluation of mislabel identification and correction methods from the participants. Individual solutions submitted by the challenge participants, even those from the same team, showed a wide range of accuracy, underscoring the importance of the benchmarking effort. Second, post-challenge collaborative efforts in validating, refining, and integrating the top-performing methods have led to an open-source product, which showed high accuracy and robustness in mislabeling identification and correction in simulated and real datasets. Third, we applied COSMO to three real datasets without prior sample-mislabeling reports and identified errors from two datasets. We further showed that error correction had a significant impact on the conclusions of the studies, thus demonstrating the potential biological impact of the tool.

There are a few limitations of our challenge design. First, due to limited data availability, one dataset was split into a training set and a test set. Because the training set and the test set are not completely independent, generalizability of the winning solutions cannot be guaranteed. Second, due to concerns on information leaking, it was unrealistic to perform repeated hold-out validation during the course of the challenge. Thus, only a single hold-out dataset was used for performance evaluation, limiting the stability of the evaluation results. These limitations were partially addressed by performing bootstrapping resampling to determine top performers and by confirming the robustness and generalizability of both the winning algorithms and the final crowdsourced product COSMO during the post-challenge development phase through repeated hold-out validation using two independent simulated datasets and application to six real multi-omic datasets. Nevertheless, when possible, future challenge designs should use multiple datasets for training and completely independent datasets for final performance evaluation. Third, to broaden challenge participation and considering the fact that some models may not be able to generate prediction probabilities (e.g., rule-based models), the crowdsourcing challenge committee had decided to require that participants submit only the final predictions. Consequently, taking into account the imbalanced class distribution, F_1 score was used for performance evaluation. Future challenge designs could require the submission of prediction scores or probabilities, which will support a more holistic evaluation using the area under the receiver-operating characteristics (AUROC) metric.

Algorithms used in COSMO were selected on the basis of the competition results. Although these algorithms outperformed others in the competition, they may not be the best solutions for solving this challenge. Moreover, because COSMO was developed primarily for proteogenomic studies involving proteomic and RNA-seq data, there are some assumptions in the current implementation that need to be considered for appropriate application to other types of multi-omic studies. There are two major steps in COSMO: one is omic data-based phenotype prediction and the other is sample matching between omic data. For the first step, COSMO is applicable to any omic data as long as the signal is sufficient for accurate phenotype prediction without labeling errors. Somatic mutations are typically reported as binary data and are typically not sufficient for phenotype prediction. However, some frequently mutated genes (e.g., TP53) might be used similarly to clinical phenotype data if they can be accurately predicted by other omic data types (e.g., RNA-seq and proteomics). For the second step, sample matching is based on correlation between omic profiles, so it is only feasible for omic data with continuous measurements and can be summarized to gene level to allow correlation analysis. For example, metabolomic data cannot be directly used in the current implementation. Moreover, the two omic profiles from the same sample must have sufficiently strong correlation without labeling errors to allow accurate sample matching. For example, the correlation between methylation and proteomics might not be sufficient for such analysis. With an unprecedented level of resolution, single-cell omics is revolutionizing biomedical research. Compared with bulk cell studies, single-cell data have unique noise properties and data sparseness. New computational algorithms are needed for identifying and correcting mislabeled samples in single-cell multi-omic studies.

Regardless of the limitations described above, COSMO showed its robust and general applicability to proteogenomic datasets with or without previous knowledge of mislabeled samples. Further analysis of these datasets showed a clear impact of sample errors in both statistical and biological aspects. Therefore, our study suggested that a sample-labeling check is an essential quality control prior to data analysis and that COSMO is a valuable tool for this task. The final product COSMO and its source code are openly available at the GitHub, thus allowing broad usage and continuous development by the global scientific community. In addition to providing a practically useful tool, we also hope that this study stimulates more research into computational methods for identifying and correcting mislabeled samples in different types of multi-omic studies.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Bing Zhang (bing.zhang@bcm.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Challenge data can be accessed at https://precision.fda.gov/challenges/5. Real-case datasets can be accessed at https://github.com/bzhanglab/COSMO. The software package COSMO is available at https://github.com/bzhanglab/COSMO. The COSMO manual is available in supplemental experimental procedures.

Challenge datasets

Merging two colon rectal cancer datasets

The transcriptomic and proteomic data of two colon rectal cancer cohorts (85 from Zhang et al.[4] and 96 from Vasaikar et al.[8]) were merged into data matrices of 181 samples. Because both studies had already been published at the time of the challenge design, we mixed samples from the two studies and reprocessed the combined data to reduce possible breach of information that participants could use as leverage. Protein quantification based on spectral counting was performed as described in a previous study[8] and mRNA quantification based on fragments per kilobase of transcript per million mapped reads (FPKM) was performed as described in the two colon rectal cancer cohorts. For both proteomic and RNA-seq data, genes with more than 50% missing values were removed, except for genes located in X or Y chromosomes, which were retained even if they were missed in more than 50% of the samples. The proteomic data were then normalized using quantile normalization followed by batch correction using ComBat,[27] whereas the RNA-seq data were normalized using the trimmed mean of M-values normalization method (TMM)[28] followed by batch correction using ComBat (Figure S4). Quality control analysis was performed using metaX[29] before and after batch correction.

Further filtering

Next, MODMatcher[13] was applied to the 181 samples to identify any ambiguously matched samples between RNA-seq and proteomic data. Among 3,882 common features from both RNA-seq and proteomics, highly correlated gene-protein pairs were used to evaluate sample similarity scores. For the purpose of challenge design, clean ground truth would be necessary for fair evaluation of the submitted solutions. Considering potential labeling errors in the original dataset, we removed 19 samples with poor sample similarity scores ($\rho < 0.5$), and the remaining 162 samples showing strong correlation between their mRNA and protein abundance were retained. Errors were then generated randomly among these samples.

Generation of mislabeling samples

Based on the previously observed patterns and rates of sample-labeling errors in various TCGA or CPTAC datasets, we introduced similar error patterns from three mislabel types: duplication, swapping, and shifting. To provide guidelines for the participant for their method development, we set the following rules for the errors. (1) We introduced labeling errors to 10% of the samples (n = 8) to proteomic data and RNA-seq data, respectively, and introduced labeling errors to 5% of the samples (n = 4) in the clinical information table. Hence, there would be a total of 20 samples with labeling errors. (2) For clinical data, we only introduced swapping between two pairs of gender-inconsistent samples so that the errors could be recognized. (3) For proteomic and RNA-seq data, all three error types, sample duplication (n = 1), sample swapping (n = 4 from 2 pairs), and sample shifting (n = 3), were generated. Duplicate samples in proteomic data were actual proteomic profiles from replicate proteomic experiments meeting the sample similarity (Figure S5, left). A duplicate sample in RNA-seq data was simulated by adding a perturbation equal to the standard deviation of each gene i as in $\text{Sample}(i)_{\text{duplicate}} > \text{Sample}(i) \pm \frac{\sigma}{\alpha}$, where σ is a standard deviation of the gene i and α is a scale factor for the σ. We changed the scale factor to generate a duplicate sample (Figure S5, middle) and, with $\alpha = 1$, correlation coefficients between simulated RNA-seq replicates and the original samples were greater than 0.9 as similarly in proteomic duplicates (Figure S5, right). (4) The swapped samples had different gender or MSI status. (5) Sample-labeling errors were not shared across different types of data (i.e., for each sample, error only happens in one type of data matrices), so that all three data types could be used to identify the sources of the errors.

Generation of training and testing datasets

From the 162 samples well matched among clinical attributes (genders and MSI), RNA-seq, and proteomic data, 80 samples were randomly selected for training and another 80 samples for testing. We then introduced mislabeling samples into proteomic, RNA-seq, and clinical information data in both the training and test sets following the above rules in the following order: (1) in the proteomic (RNA-seq) matrices, one sample was randomly selected and then replaced with the replicate of another sample in the remaining set; (2) from the samples without replicates, two pairs of samples were randomly selected and their sample labels were swapped; (3) in the remaining samples, three samples were randomly selected and their labels were shifted (A to B, B to C, and C to D). We repeated these steps 50 times to generate random pairs of training and testing datasets.

Selection of the challenge problem set

Our baseline method starts with using molecular data (RNA-seq, protein) as features to predict patient gender and MSI status. Here we only used sex genes for predicting the gender while using all available genes to predict MSI status. We trained XGBoost models with AUROC as an evaluation metric. Hyperparameters were determined by a 3-fold cross-validation grid search. For each model, we define the prediction error as the absolute difference between the provided binary value and the predicted probability of the sample being positive class. Accordingly, each sample now has four prediction error scores: $\delta_{\text{rna, gender}}$, $\delta_{\text{rna, msi}}$, $\delta_{\text{pro, gender}}$, and $\delta_{\text{pro, msi}}$. For sub-challenge 1, a sample is considered a mismatch between clinical and protein profiling data only when both $\delta_{\text{pro, gender}} > 0.5$ and $\delta_{\text{pro, msi}} > 0.5$ are true. For each data type (RNA-seq, protein), we further sum the prediction errors of both phenotypes to obtain two scores: *rna_score* and *pro_score*. Finally, the *clin_score* is defined as the sum of *rna_score* and *pro_score*. For each sample, *clin_score* indicates overall how well the provided molecular data can predict its clinical phenotypes. We denote samples with questionable clinical data as $S_c = \{S_i \mid clin\_score < 3\}$. The rationale behind this is: if the predictions are simply random, the *clin_score* will be 2. At the other extreme, if the predictions are all perfect, the *clin_score* is 4. We think a score of 3 is a reasonable cutoff value. Next, we perform protein-RNA-seq data correlation analysis to evaluate the mismatch between these two data types. For each gene, we calculate the Spearman correlation coefficient between RNA and protein data after scaling. The top 200 most correlated genes g_{top} are selected. With the selected g_{top}, we compute Spearman correlation between RNA-seq and protein data for each sample and get $\rho = \{\rho_{1,\ldots,\rho_{80}}\}$. We then define the outlier threshold as $\theta = \text{median}(\rho) - 2 * \text{MAD}(\rho)$, where MAD is the median absolute deviation. Any sample with correlation coefficient less than the threshold is labeled as questionable sample with unmatched protein-RNA-seq data, i.e., $S_{pr} = \{S_i | \rho_i < \theta\}$. From the set of S_{pr}, we further identify samples with questionable RNA-seq data and with questionable protein data as $S_r = \{rna\_score_i < pro\_score_i \mid S_i \in S_{pr}\}$ and $S_p = \{rna\_score_i \geq pro\_score_i \mid S_i \in S_{pr}\}$, respectively. In our baseline analysis, we do not intend to correct the labels of sample in *Sc* and instead set the corrected label as -1. However, for samples in both S_r and S_p, we employ the cross-data type correlation analysis to assign the corrected label. Specifically, for each RNA-seq and protein sample pair (S_i, S_j), we compute its Spearman correlation coefficient ρ_{ij} with the top genes g_{top}. For each sample i in S_r, we set its corrected label to the label of the protein sample with which sample i has the largest correlation coefficient, i.e., to the label of sample k, where $k = \text{argmax}_j \rho_{ij}$. The same criteria also apply to each sample in S_p. We apply the baseline pipeline to the 50 randomly generated training and test dataset pairs. For each pair, we obtain the average F_1 score of sub-challenge 1 and three sub-challenge 2 scores. We then select the pair with the median average score (F_1 score = 0.68, Figure S1) as the final dataset for the competition.

Evaluations of the challenge submission

Measurement of F_1 scores

Each submission was evaluated by F_1 score, the harmonic mean of precision and recall, as $F_1 = 2 \times \left(\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}\right)$. The submitted data matrix of 80 samples in the testing dataset were compared with the answer sheet. For the sub-challenge 1, the F_1 score was measured directly but for the sub-challenge 2, we evaluated the model performance at three different levels. (1) How well the model predicted mislabeling at the sample level: if any of the predicted labels

does not match the original sample label, it is considered a mislabel at this level. (2) How the model identified the source of errors among three types of data: label prediction from clinical and omic profile data are compared with the original labels. A prediction that correctly identifies a mislabel, but not necessarily correctly rectifies it, will be considered as a true positive. (3) How well the model corrected the errors by matching accurate samples: only when a corrected label matches the true sample label will it be considered a true positive. The F_1 scores at these three levels were then averaged for the final score.

Determination of top performers

Confidence interval (CI) of F_1 score of each submission was calculated by performing bootstrapping resampling.[30] First, a sub-set of 60 out of 80 samples were randomly selected and the F_1 score was measured based on the 60 samples. The resampling was iterated 100 times to generate mean and standard deviation from bootstrap estimate distribution of F_1 score. Next, the 95% CI was measured as $\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$. Multiple submissions from the same group were then averaged for final determination of challenge winners.

Simulated datasets with random types/events of errors

For the validation datasets, we used two independent cohorts. From the 162 colon dataset, 100 random samples were randomly selected and the errors were introduced, but the number of errors was not fixed to mimic real-case scenarios. We randomly introduced three types of errors in up to 28 samples out of 100. This procedure was iterated 50 times to generate random distribution of different types of errors. In addition, we also used 110 CCRCC tumor samples to generate 50 random error-containing datasets including RNA-seq and global proteomic data with associated gender information.

Development of COSMO

The COSMO algorithm works as follows: data preprocessing, pairwise alignment, clinical attribute prediction, and label correction. We take RNA-seq and proteomic data as an example in the following method description. However, the application of COSMO is not limited to specific platforms or data types and can be equally applied to other types of gene-centric datasets, such as gene-level Riboseq or CNV data.

Data preprocessing

The genes in both RNA-seq and proteomic data are annotated with chromosome information and categorized into sex-linked genes or autosomal genes. The annotation determines how missing values in the data are handled. Missing values of sex-linked genes are replaced with 0, as these genes are assumed to be either absent (i.e., the absence of the Y chromosome in female) or repressed (i.e., X chromosome inactivation in male). For autosomal genes, genes that have missing values in >50% of samples are removed. For the remaining genes, the missing values are either removed or imputed via RF missing data imputation. Missing data imputation requires a noticeable time, and the decision to do imputation depends on the portion of genes with missing values. In our work, if the removal of missing values will result in a loss of >30% of the data, the missing values are imputed.

Pairwise alignment

Mislabeled samples would constitute noise, and a prediction model trained using the entire dataset will result in a low prediction performance. Thus, before training prediction models, we perform pairwise alignment to determine the mislabeled samples to exclude them in model training. We exploit the parallel nature of different omic data and computed correlation signals to pair RNA-seq and proteomic samples. Corresponding samples are samples with the same label, i.e., RNA-seq sample (r_7) is corresponding with the proteomic sample (p_7), as both have the same label (7), indicating that both of them belong to a particular patient where index = 7. If there is no mislabeling, the corresponding samples should have the highest correlation signal with each other and be paired together.

The correlation signal is computed from the omic data matrix. For every autosomal gene that exists in both RNA-seq and proteomic data, g, we compute its inter-omic correlation using Equation (1).

$$Cor_g = \mathrm{cor}(R_{*g}, P_{*g}), \text{where } g \in OG. \quad \text{(Equation 1)}$$

Cor_g is the inter-omic correlation of gene g, R_g is the vector of mRNA expression values of gene g across samples, while P_g is the vector of protein expression values of protein g across the same samples. OG is the set of all overlapping genes present in both RNA-seq and proteomic data. Genes with inter-omic correlation >0.5 are extracted. The expression values of the extracted genes are used to compute the inter-samples correlation. In inter-samples correlation matrix, C indicates the correlation of any RNA-seq samples with any proteomic sample, with a dimension of $N \times N$.

$$C_{ij} = \mathrm{cor}(R_i, P_j). \quad \text{(Equation 2)}$$

C_{ij} is the inter-samples correlation of RNA-seq sample i with proteomic sample j. R_i is the vector of mRNA expression values of sample i across the extracted genes, while P_j is the vector of protein expression values of sample j across the same genes.

The correlation matrix contains only the degree of association between any pair of samples, regardless of the association with other samples. We derived a probability matrix, PM using Equations (3), (4), and (5). The probability matrix PM incorporates the degree of association among other samples and ranges within 0–1, and scales the range where every RNA sample has ~1 probability distributed to every PRO sample and vice versa. PM_{ij} indicates the probability of RNA-seq sample i match to proteomic sample j. The pair of matching samples will have the highest probability with each other.

$$C\_rna_{i*} = \mathrm{softmax}(\mathrm{standardize}(C_{i*})), \quad \text{(Equation 3)}$$

$$C\_pro_{*j} = \mathrm{softmax}(\mathrm{standardize}(C_{*j})), \quad \text{(Equation 4)}$$

$$PM = \sqrt{Crna \times Cpro}. \quad \text{(Equation 5)}$$

The probability matrix PM is used as the preferential ranking for stable matching algorithm, with the highest probability as the most preferred candidate and the lowest the least preferred. The stable matching algorithm generates N pairs of RNA-seq and proteomic samples, where N = number of tissue sample. Every samples pair has a matching score for every pairing, which is the sum of the preferential rank of RNA-seq sample toward proteomic sample and vice versa. An ideal pairing should have a matching score of 2, indicating that both RNA-seq and proteomic samples have the strongest correlation signals with each other.

Corresponding samples that are paired together are considered correctly labeled, and these samples are called matching samples. Samples that are not matched with its corresponding samples are mislabeled. Matching samples with a matching score >log(N), where N = number of tissue samples, are also considered mislabeled. This is because the stable matching algorithm will pair any RNA-seq sample with exactly one proteomic sample, thus samples that are left out (due to duplication) will be paired despite having very low correlation signals with each other.

The correlation signal is computed again in the second iteration, when the correlated genes are extracted using only matching samples to obtain a more accurate correlation signal and, thus, a new set of matching samples with higher confidence. Matching samples will be inspected for clinical swapping cases and used to train the prediction models (see below). On the other hand, mismatched samples are retained for label correction (see subsequent section on label correction).

Clinical attribute prediction

Two methods from the top-performing teams were improved from the versions used in the challenge and then integrated for clinical attribute prediction.

Method 1. Matching samples are samples with no RNA-seq or proteomic mislabeling, since corresponding RNA-seq and proteomic samples are paired with each other. However, this does not preclude the occurrence of clinical data mislabeling. Clinical swapping cases constitute noise in training prediction models, albeit with a low frequency (~5%).[13] Thus, clinical attribute prediction is performed in a two-iterations manner. The first iteration is to determine those clinical swapping cases and exclude them from model training in the second iteration.

Every sample has a clinical profile with clinical attributes labeled manually. If it is labeled correctly, its clinical profile should be consistent with its omic profile. An omic profile is a profile with clinical attributes predicted from RNA-seq and proteomic data. We used two clinical attributes (MSI and

gender) in the colorectal dataset and one clinical attribute (gender) in the kidney dataset.

In the first iteration, the clinical attribute of RNA-seq and proteomic data are predicted using 5-fold cross-validation. In each fold, four parts of the matching samples are used for training while the remaining part is used for testing. Since these samples are matching samples (i.e., both RNA-seq and proteomic data should have the same clinical attribute), the predicted probability of RNA-seq and proteomics are averaged to form an omic profile.

We compute the error rate of every sample, which is the difference between its clinical profile and omic profile. Samples that have an error rate >0.5 are considered as potential mislabeled samples and are filtered out. The stable matching algorithm is then deployed to pair the omic profile and clinical profile of the filtered samples, using error rate as the preferential ranking (the lowest being the most preferred candidate and the highest the least preferred). Due to false prediction, it is possible that filtered samples are in odd numbers, but the stable matching algorithm is robust in this issue where the remaining sample would be paired with itself. Those samples whose clinical profile does not pair with their own omic profile will be determined as potential clinical swapping cases.

In the second iteration, potential clinical swapping cases are removed from matching samples and the remaining matching samples are used to train the prediction models. We use weighted LR with L1 regularization for training and different models are trained for different clinical attributes from different omic data. The trained models are used to predict clinical attributes of every sample, including the potential clinical swapping cases and mismatch samples. Using the newly predicted attribute, the matching samples are then inspected again for clinical swapping cases by the same process in the first iteration (building the omic profile, determining the error rate, filtering samples with error rate >0.35, and feeding into the stable matching algorithm). Determined clinical swapping cases have their label corrected.

Method 2. For each dataset, highly correlated features (Pearson correlation coefficient >0.9) were first removed. A classifier was built for each clinical attribute. The classifier is an ensemble of LR with L1 and L2 regularization, respectively, enabling automatic feature selection in the fitting process. Hyperparameters for the predictive models are chosen through cross-validation. The modeling was repeated 100 times and the predicted probabilities for each sample were then averaged to generate a final probability for each sample. If multiple clinical attributes were provided, the predicted probabilities for different clinical attributes were then combined together to obtain a multi-class label mismatch classifier.

Label correction

The prediction models from both methods are integrated and are also used to predict the clinical attributes of mismatched samples. We devised a correction algorithm that utilized the pairwise alignment and predicted attributes. For clinical swapping cases, the stable matching algorithm is used to pair the clinical profile with the omic profile before correcting the label. For RNA-seq and proteomic mislabeling, the algorithm determines which type of omics are mislabeled by comparing the error rate between RNA-seq mislabeling and proteomic mislabeling. There are different mislabeling error types (swapping, duplication, and shifting), and the exact mechanism of correction is different for each of them.

Swapping cases are the most easily identified mislabeling type. Two different patients will have both their RNA-seq and proteomic samples matched with each other. To determine whether it is RNA-seq or proteomic swapping, the predicted attributes are inspected and the prediction probability is used to compute the error rate. Considering the error rate after swapping RNA-seq samples and after swapping proteomic samples, the one that gives a lower error rate will have their labels corrected. In other words, if swapping RNA-seq samples results in a lower error rate than swapping proteomic samples, it is considered as RNA-seq swapping and the labels of RNA-seq samples are corrected.

The stable matching algorithm pairs one RNA-seq sample with exactly one proteomic sample. This complicates the identification of duplication cases, as a duplicated sample will not pair with its matching sample. Hence, the identification of duplication cases relies on the matching score. The duplication sample is always spuriously paired with another sample and will have a matching score higher than a threshold, $\log_2(N)$. We use $\log_2(N)$ as the threshold to allow higher flexibility of spurious correlation in a dataset with higher sample numbers. Consider a case where a sample pair with a matching score higher than $\log_2(N)$ is suspected to be a duplication. Here, the most preferred candidate for RNA-seq sample and proteomic sample is further inspected. This leads to another two possible sample pairs, and the next step is to determine whether it is RNA-seq or proteomic duplication. The derived probabilities of these two potential sample pairs are compared, and the one with higher probability will have its label corrected.

Shifting cases always start with a duplication event. Before correcting the label, one has to identify the shifting chain. The shifting chain starts with a duplicated sample, which is identified as described in the previous paragraph. The chain is identified by iteratively inspecting the sample pair of the last sample in the chain until the chain reaches a sample pair with a score higher than $\log_2(N)$. This is due to the nature of a stable matching algorithm, pairing one RNA-seq sample with exactly one proteomic sample. The samples with no matching samples are left out and thus are spuriously paired but with a high matching score. After the shifting chain is identified, the next step is to determine whether it is RNA-seq or proteomic shifting by classification probability. Considering the error rate after shifting RNA-seq samples and after shifting proteomic samples, the one that gives a lower error rate will have its labels corrected. In other words, if shifting RNA-seq samples results in a lower error rate than shifting proteomic samples, it is considered RNA-seq shifting, and the labels of RNA-seq samples are corrected.

Implementation of COSMO using Nextflow and Docker

The COSMO workflow was implemented using Nextflow[31] and Docker. Specifically, all the dependencies were containerized as a single Docker image. Different components of COSMO were integrated using Nextflow. The input files required by COSMO include protein expression file and gene expression file at RNA level, as well as a sample annotation file containing clinical information of samples. The source code of COSMO is available at https://github.com/bzhanglab/COSMO.

Real-case datasets with mislabeling samples

Six independent previously published datasets including labeling errors were further used to evaluate the performance of COSMO (Table S3). For the preQC CPTAC LUAD dataset,[21] we used 107 tumor samples in RNA-seq, global proteomic, and CNV profiles as well as gender information. For the preQC CPTAC CCRCC dataset,[7] we collected 77 tumor tissues in RNA-seq, global proteomic, and CNV profiles with clinical information. All the errors in the CPTAC LUAD and CCRCC datasets were corrected after initial observation, and currently released data are error free. For TCGA BRCA dataset,[6] we downloaded 521 tumor samples in microarray, RNA-seq, and CNV from the TCGA data portal (https://portal.gdc.cancer.gov/). For the CCLE data,[24] we downloaded gene expression, global proteomic, and copy-number profiles from CCLE the dataset (https://portals.broadinstitute.org/ccle/data), and selected 371 samples having all three types of data. For the two non-cancerous proteogenomic datasets, Battle et al.[22] and Chick et al.,[23] we downloaded their published data as instructed in their publications. All downloaded omic data were arranged in the same format of rows (genes) and columns (samples) to be used as input of COSMO. CNV data were downloaded directly from the original studies, and data preprocessing was diverse in each cohort. For CPTAC CNV, whole-genome and exome sequencing data were used to estimate circular binary segmentation means using an algorithm called CNVEX.[7,21] The TCGA BRAC study used GISTIC2[6] while the CCLE study used ABSOLUTE for CNV data analysis.[24]

DEPs were identified based on t test (FDR < 0.05) between two tumor groups separated by gender or immune sub-types. Functional enrichment test of the DEPs was performed by Fisher's exact test (FET) against 50 HALLMARK pathways,[32] and significant pathways were determined by FET with FDR < 0.05.

CONSORTIA

The FDA-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge Consortium: Francisco Azuaje, Ruth Bandler, Ancha Baranova, Ranjan Kumar Barman, Christophe Battail, Katharina Baum, Omid Bazgir, Adam Berger, Emily Boja, Tomáš Brůna, Anders Carlsson, Wencai Cao, Antonio

Cappuccio, Alexis Carter, Hong Chen, Libin Chen, Feixiong Cheng, Yeshwant Chillakuru, Junseok Choe, John Chung, Ana Rita Colaço, Tiange Cui, MengHua Deng, Saugato Rahman Dhruba, John Didion, Patrik Edén, Enhao Fang, Hanying Feng, Slim Fourati, Xiao Dan Gan, Xijin Ge, Souparno Ghosh, Bruno Giotti, Xiangkun Gu, Marouen Ben Guebila, Baosen Guo, Jing Guo, Alexander Harms, Majda Haznadar, Jing Hu, Dan Huang, Maven Hyun, Yingjie Ji, Xiaoqing Jiang, Elaine Johanson, Jaewoo Kang, Tony Kaoma, SoonJye Kho, Keonwoo Kim, Sang-Yoon Kim, Sunkyu Kim, Sindhu Kubendran, Rajnish Kumar, Rintu Kutum, Alden Leung, K.S. Leung, Dehua Li, Björn Linse, Xiangrong Liu, Singer Ma, Weiping Ma, Ezekiel J. Maier, Matthias Mann, Arnaud Muller, Vijayaraj Nagarajan, Petr Nazarov, Anne-laura Bach Nielsen, Thin Nguyen, Mattias Ohlsson, Ranadip Pal, Renke Pan, Arjun Panda, Dana Pascovici, Samuel Payne, Carsten Peterson, Sarah Prezek, Pingfa Qin, Raziur Rahman, Michael Raymer, Michael Robben, Henry Rodriguez, Alberto Santos, Daniel Schlauch, Lawrence Segbehoe, Linghao Shen, Zhiao Shi, Li Ka Shing, Gustavo Stolovitzky, Nelson Tang, Zivana Tezak, Boris Veytsman, Raymond Wan, Cankun Wang, Pei Wang, Samuel Wang, Bo Wen, Jemma X. Wu, Longxin Wu, Canqiang Xu, Wenxian Yang, Runan Yao, Seungyeul Yoo, Rongshan Yu, Elena Zaslavsky, Xiangxiang Zeng, Bing Zhang, Yue Zhang, Xiaowei Zhan, Hancheng Zheng, Lisheng Zhou, Wengang Zhou, Tao Zhou, Jun Zhu, Yun Zuo.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100245.

ACKNOWLEDGMENTS

This study was supported by the National Cancer Institute CPTAC awards U24CA210954 and U24CA210993, the Cancer Prevention & Research Institutes of Texas (CPRIT) award RR160027, and funding from the McNair Medical Institute at The Robert and Janice McNair Foundation. B.Z. is a CPRIT Scholar in Cancer Research and a McNair scholar. S.K. acknowledges the funding provided by Wright State University. We thank all organization committee members and contestants of the FDA-CPTAC Multi-omics Enabled Sample Mislabeling Correction Challenge (Document S2) for their contribution to this study

AUTHOR CONTRIBUTIONS

Conceptualization, B.Z., P.W., and E.B.; Methodology, S.Y., Z.S., B.W., S.K., R.P., H.F., H.C., A.C., P.E., W.M., M.R., and J.Z.; Software, S.Y., Z.S., B.W., S.K., R.P., H.F., and A.C.; Formal analysis, S.Y., Z.S., B.W., S.K., R.P., H.F., A.C., P.W., and B.Z.; Investigation, S.Y., Z.S., B.W., S.K., R.P., H.F., A.C., P.W., and B.Z.; Writing, S.Y., Z.S., B.W., S.K., E.B., P.W., and B.Z.; Supervision, J.Z., P.W., and B.Z.; Project administration, E.J.M., Z.T., E.J., D.H., H.R., E.B., P.W., and B.Z. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

S.Y. and J.Z. are employees of Sema4, a for-profit organization that promotes a healthcare through information-driven insights. R.P., H.F., and H.C. are employees of Sentieon Inc. A.C. is an employee of Bionamic AB. The other authors declare no competing interests.

Received: December 23, 2020
Revised: January 27, 2021
Accepted: March 31, 2021
Published: May 7, 2021

REFERENCES

1. Nilsson, T., Mann, M., Aebersold, R., Yates, J.R., 3rd, Bairoch, A., and Bergeron, J.J. (2010). Mass spectrometry in high-throughput proteomics: ready for the big time. Nat. Methods *7*, 681–685. https://doi.org/10.1038/nmeth0910-681.

2. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.

3. Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. Nat. Rev. Genet. *20*, 631–656. https://doi.org/10.1038/s41576-019-0150-2.

4. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387. https://doi.org/10.1038/nature13438.

5. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.L., Tokheim, C., et al. (2018). Perspective on oncogenic processes at the end of the beginning of cancer genomics. Cell *173*, 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033.

6. Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.

7. Clark, D.J., Dhanasekaran, S.M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.M., Chang, H.Y., et al. (2020). Integrated proteogenomic characterization of clear cell renal cell carcinoma. Cell *180*, 207. https://doi.org/10.1016/j.cell.2019.12.026.

8. Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell *177*, 1035–1049.e19. https://doi.org/10.1016/j.cell.2019.03.030.

9. Dou, Y., Kawaler, E.A., Cui Zhou, D., Gritsenko, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Satpathy, S., et al. (2020). Proteogenomic characterization of endometrial carcinoma. Cell *180*, 729–748.e6. https://doi.org/10.1016/j.cell.2020.01.026.

10. Astion, M.L., Shojania, K.G., Hamill, T.R., Kim, S., and Ng, V.L. (2003). Classifying laboratory incident reports to identify problems that jeopardize patient safety. Am. J. Clin. Pathol. *120*, 18–26. https://doi.org/10.1309/8EXC-CM6Y-R1TH-UBAF.

11. College of American, P., Valenstein, P.N., Raab, S.S., and Walsh, M.K. (2006). Identification errors involving clinical laboratories: a College of American Pathologists Q-Probes study of patient and specimen identification errors at 120 institutions. Arch. Pathol. Lab. Med. *130*, 1106–1113. https://doi.org/10.1043/1543-2165(2006)130[1106:IEICL]2.0.CO;2.

12. Toker, L., Feng, M., and Pavlidis, P. (2016). Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. F1000Res. *5*, 2103. https://doi.org/10.12688/f1000research.9471.2.

13. Yoo, S., Huang, T., Campbell, J.D., Lee, E., Tu, Z., Geraci, M.W., Powell, C.A., Schadt, E.E., Spira, A., and Zhu, J. (2014). MODMatcher: multi-omics data matcher for integrative genomic analysis. Plos Comput. Biol. *10*, e1003790. https://doi.org/10.1371/journal.pcbi.1003790.

14. Lee, E., Yoo, S., Wang, W., Tu, Z., and Zhu, J. (2019). A probabilistic multi-omics data matching method for detecting sample errors in integrative analysis. Gigascience *8*. https://doi.org/10.1093/gigascience/giz080.

15. Lee, S., Lee, S., Ouellette, S., Park, W.Y., Lee, E.A., and Park, P.J. (2017). NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. Nucleic Acids Res. *45*, e103. https://doi.org/10.1093/nar/gkx193.

16. Simpson, J.B. (2001). A unique approach for reducing specimen labeling errors: combining marketing techniques with performance improvement. Clin. Leadersh. Manag. Rev. *15*, 401–405.

17. Huang, J., Chen, J., Lathrop, M., and Liang, L. (2013). A tool for RNA sequencing sample identity check. Bioinformatics *29*, 1463–1464. https://doi.org/10.1093/bioinformatics/btt155.

18. Javed, N., Farjoun, Y., Fennell, T.J., Epstein, C.B., Bernstein, B.E., and Shoresh, N. (2020). Detecting sample swaps in diverse NGS data types using linkage disequilibrium. Nat. Commun. *11*, 3697. https://doi.org/10.1038/s41467-020-17453-5.

19. Lohr, M., Hellwig, B., Edlund, K., Mattsson, J.S., Botling, J., Schmidt, M., Hengstler, J.G., Micke, P., and Rahnenfuhrer, J. (2015). Identification of sample annotation errors in gene expression datasets. Arch. Toxicol. *89*, 2265–2272. https://doi.org/10.1007/s00204-015-1632-4.

20. Boja, E., Tezak, Z., Zhang, B., Wang, P., Johanson, E., Hinton, D., and Rodriguez, H. (2018). Right data for right patient—a precision FDA NCI-CPTAC Multi-omics Mislabeling Challenge. Nat. Med. *24*, 1301–1302. https://doi.org/10.1038/s41591-018-0180-x.

21. Gillette, M.A., Satpathy, S., Cao, S., Dhanasekaran, S.M., Vasaikar, S.V., Krug, K., Petralia, F., Li, Y., Liang, W.W., Reva, B., et al. (2020). Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell *182*, 200–225.e5. https://doi.org/10.1016/j.cell.2020.06.013.

22. Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. Science *347*, 664–667. https://doi.org/10.1126/science.1260793.

23. Chick, J.M., Munger, S.C., Simecek, P., Huttlin, E.L., Choi, K., Gatti, D.M., Raghupathy, N., Svenson, K.L., Churchill, G.A., and Gygi, S.P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. Nature *534*, 500–505. https://doi.org/10.1038/nature18270.

24. Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the cancer cell line Encyclopedia. Nature *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.

25. Feldman, D.E., Chen, C., Punj, V., and Machida, K. (2013). The TBC1D15 oncoprotein controls stem cell self-renewal through destabilization of the Numb-p53 complex. PLoS One *8*, e57312. https://doi.org/10.1371/journal.pone.0057312.

26. Yang, Y., Li, M., Yan, Y., Zhang, J., Sun, K., Qu, J.K., Wang, J.S., and Duan, X.Y. (2015). Expression of RAP1B is associated with poor prognosis and promotes an aggressive phenotype in gastric cancer. Oncol. Rep. *34*, 2385–2394. https://doi.org/10.3892/or.2015.4234.

27. Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118–127. https://doi.org/10.1093/biostatistics/kxj037.

28. Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25. https://doi.org/10.1186/gb-2010-11-3-r25.

29. Wen, B., Mei, Z., Zeng, C., and Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. BMC Bioinformatics *18*, 183. https://doi.org/10.1186/s12859-017-1579-y.

30. Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. U S A *93*, 7085–7090. https://doi.org/10.1073/pnas.93.14.7085.

31. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319. https://doi.org/10.1038/nbt.3820.

32. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.