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.
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 Eden,8 Weiping Ma,1,2 Michael Raymer,10 Ezekiel J. Maier,10 Zivana Tezak,11 Elaine Johanson,11 Denise Hinton,11 Henry Rodriguez,14 Jun Zhu,1,2,3 Emily Boja,14 Pei Wang,1,2,* and Bing Zhang4,5,16,*

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

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. 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
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.\(^1\)\(^-\)\(^6\) 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.\(^1\)\(^-\)\(^9\) 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 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 \textit{XIST} or \textit{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,\(^13\)\(^,\)\(^14\)\(^,\)\(^16\) 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.\(^7\) 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.
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 F1 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 F1 score of 0.75, but generalizability of this approach is uncertain because it was used by only one team.

For sub-challenge 1 the F1 scores ranged from 0 to 0.83, 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 F1 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
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 F1 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 F1 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 F1 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 F1 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 F1 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 F1 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...
high accuracies (F1 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), the preQC CPTAC kidney cancer dataset (preQC CPTAC CCRCC), and the TCGA breast cancer dataset (TCGA BRCA) (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-Protomics, RNA-seq-CNV (copy-number variation), and Protomics-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. 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) 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, 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, 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 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...
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_CHE CKPINT, E2F_TARGETs, MYC_TARGETS_V1, and MYC_TAR-
GETS_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. 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 the COSMO-corrected data, and stronger associations were observed for other immune response-related pathways such as INTERFERON_GAMMA_RESPONSE and
ALLOGRAFT_REACTION.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

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...
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-performing solutions 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, F1 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**

**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.1 and 96 from Vasaikar et al.)13 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 study14 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,15 whereas the RNA-seq data were normalized using the trimmed mean of M-values normalization method (TMM)26 followed by batch correction using ComBat (Figure S4). Quality control analysis was performed using metaX16 before and after batch correction.

**Further filtering**
Next, MODMatcher17 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 (p < 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 Sample(i)perturbed = Sample(i) + s · rand(), where s is a standard deviation of the gene i and a is a scale factor for the s. We changed the scale factor to generate a duplicate sample (Figure S5, middle) and, with a = 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 predicted value and the predicted probability of the sample being positive class. Accordingly, each sample now has four prediction error scores: δg, δmsi, δg, δmsi. For sub-challenge 1, a sample is considered a mismatch between clinical and protein profiling data only when both δg, δmsi > 0.5 and δg, δmsi > 0.5 are true. For each data type (RNA-seq, protein), we further sum the prediction errors of both phenotypes to obtain two scores: ma_score and pro_score. Finally, the clin_score is defined as the sum of ma_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 = {Sr | 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 gtop are selected. With the selected gtop, we compute Spearman correlation between RNA-seq and protein data for each sample and get r = {ρg, ρg}. We then define the outlier threshold as θ = median(ρ) − 2 · MAD(ρ), 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., Sρ = {Sr | ρg < θ}. From the set Sρ, we further identify samples with questionable RNA-seq data and with questionable protein data as Sg = {ma_score < ρg | Sρ ∩ Sρ} and Sp = {pro_score < ρg | Sρ ∩ Sρ}, 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, in samples in both Sρ and Sp, we employ the cross-data type correlation analysis to assign the corrected label. Specifically, for each RNA-seq and protein sample pair (Sr, Sg), we compute its Spearman correlation coefficient ρ with the top genes gtop. For each sample i in Sρ, 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 = argmax(ρg). The same criteria also apply to each sample in Sp. We apply the baseline pipeline to the 50 randomly generated training and test dataset pairs. For each pair, we obtain the average F1_score of sub-challenge 1 and three sub-challenge 2 scores. We then select the pair with the median average score (F1_score = 0.68, Figure S1) as the final dataset for the competition.

**Evaluations of the challenge submission**
**Measurement of F1_scores**
Each submission was evaluated by F1_score, the harmonic mean of precision and recall, as F1 = 2 × (precision × recall) / (precision + recall). The submitted data matrix of 80 samples in the testing dataset were compared with the answer sheet. For the sub-challenge 1, the F1_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 correctly rectifies it, will be considered as a true positive. The F1 scores at these three levels were then averaged for the final score.

**Determination of top performers**

Confidence interval (CI) of F1 score of each submission was calculated by performing bootstrapping resampling. First, a sub-set of 60 out of 80 samples were randomly selected and the F1 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 F1 score. Next, the 95% CI was measured as 

\[
\left[ \bar{x} - 1.96 \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right].
\]

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 CRCRC 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 \(R_i\) is corresponding with the proteomic sample \(P_j\), as both have the same label \(l\), indicating that both of them belong to a particular patient where index \(i \times j = 7\). If there is no labeling bias, 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).

\[
Cor_g = \text{cor}(R_g, P_g), \quad \text{where} \quad g \in \text{OG}.
\]

\(Cor_g\) 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 \(N \times N\).

\[
C = \text{cor}(R_i, P_j).
\]

\(C\) 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. \(PM\) 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.

\[
C_{\text{RNA}} = \text{softmax}(\text{standardize}(C_{ij})),
\]

(3)

\[
C_{\text{PRO}} = \text{softmax}(\text{standardize}(C_{ji})),
\]

(4)

\[
PM = \sqrt{C_{\text{RNA}}} \times C_{\text{PRO}}.
\]

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%). 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
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. Hyper-parameters 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 both have 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 samples are usually paired with another duplicated 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 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, \(^{23}\) we used 107 tumor samples in RNA-seq, global proteomic, and CNV profiles as well as gender information. For the preQC CPTAC CRCRC 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 CRCRC datasets were corrected after initial observation, and currently released data are error free. For TCGA BRCA dataset, \(^{24}\) 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, \(^{25}\) we downloaded gene expression, global proteomic, and copy-number profiles directly from CCLE (https://portals.broadinstitute.org/ccle/data), and selected 371 samples having all three types of data. For the two noncancerous proteogenomic datasets, Battie et al. \(^{26}\) and Chick et al. \(^{27}\), 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. \(^{27}\) The TCGA BRAC study used GISTIC2 \(^{26}\) while the CCLE study used ABSOLUTE for CNV data analysis. \(^{28}\) 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. \(^{29}\) 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 Dhruvs, John Didion, Patrik Edén, Enhao Fang, Hanying Feng, Slim Fourati, Xiao Dan Gan, Xijin Ge, Souphinx Ghosh, Bruno Giotti, Xiangkun Gu, Marouen Ben Guebila, Baoesen Guo, Jing Guo, Alexander Harms, Majda Haznadar, Jing Hu, Dan Huang, Maven Hyun, Yingjie Ji, Xiaoqing Jiang, Elaine Johansson, Jae-woo 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, Bjorn Linse, Xiangrong Liu, Singer Ma, Weiping Ma, Ezekiel J. Maier, Matthias Mann, Arnaud Muller, Vijayaraj Nagarajan, Petr Nazarov, Annelaure Bach Nielsen, Thin Nguyen, Mattias Ohlsson, Ranadip Pal, Renke Pan, Arjun Panda, Dana Pascovic, Samuel Payne, Carsten Petersen, Sarah Prezek, Pingfa Qin, Raziur Rahman, Michael Raychaux, Michael Robben, Henry Rodriguez, Alberto Santos, Daniel Schlauch, Lawrence Segbehaye, Lihong Shen, Zhaoyi 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, Can-qiang Xu, Wenxian Yang, Runan Yao, Seungyeol 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. The other authors declare no competing interests.

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.

REFERENCES

1. Nilsson, T., Mann, M., Aebersold, R., Yates, J.R., 3rd, Bairoeh, 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., Shaddock, 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., Thorisson, 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 Leprevest, F., Reva, B., Lih, T.M., Chiang, 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., Kawai, E.A., Cui Zhou, D., Griskeno, M.A., Huang, C., Blumenberg, L., Karpova, A., Petyuk, V.A., Savage, S.R., Sapath, 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., Shojaian, 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:IEDC2.0.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: multimomics 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 multimomics 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., Geistland, 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. DiTommaso, 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.
Supplemental information

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

Seungyeul Yoo, Zhiao Shi, Bo Wen, SoonJye Kho, Renke Pan, Hanying Feng, Hong Chen, Anders Carlsson, Patrik Edén, Weiping Ma, Michael Raymer, Ezekiel J. Maier, Zivana Tezak, Elaine Johanson, Denise Hinton, Henry Rodriguez, Jun Zhu, Emily Boja, Pei Wang, and Bing Zhang
Supplementary Figure 1. Selection of the challenge dataset among 50 simulated data sets. We applied our baseline method (Methods) on the 50 simulated data sets with fixed types and numbers of errors from the colon cancer data set. Then, the F₁ score was measured for the clinical information (in red) for the sub-challenge 1, and for Sample (in light green), Data (in green), Correction (in blue), and Average (in red), respectively, for the sub-challenge 2. Finally, we selected the dataset #17 which corresponded to the median value of Average F₁ score = 0.6836.
Supplementary Figure 2. The associations between error rate and F-scores at Sample, Data, and Correction levels, respectively. The Pearson’s correlation coefficients with corresponding p-values were shown.
Supplementary Figure 3. Biological impact of labeling error correction using COSMO in CPTAC CCRCC dataset.
Comparing correlation coefficients all 5035 gene-protein pairs before and after correction. The Pearson correlation p-values were adjusted as Benjamini-Hochberg adjusted p-values (FDR) and then log$_{10}$ transformed. Red line indicates y=x. B. Difference of correlation strengths of 54 gene-protein pairs significant only after error correction.
Supplementary Figure 4. Normalization and batch effect correction of Proteomics and RNAseq data from two independent colon cancer data sets. A. PCA score plot of raw proteomics data. B. PCA score plot of proteomics data with normalization and batch correction. D. PCA score plot of raw RNAseq data. D. PCA score plot of RNAseq data after normalization and batch correction.
**Supplementary Figure 5.** Generation of duplicates in proteomics and RNAseq data for the FDAchallenge. **A.** Histogram of sample correlation (all x all) in proteomics data. Experimental duplicates of sample #25 and #32 showed relatively high correlation with their duplicated samples in proteomics dataset. This serves as reference to generate duplicates of RNAseq data. **B.** Simulation of the scale factor $\alpha$ to control deviation expression from the original sample. The duplicated samples generated with different $\alpha$ were used to measure sample similarity with proteomics data. Score difference was measured by $\text{score difference} = \text{Score}\_self - \text{mean(Score\_others)}$. We chose $\alpha = 1$ which corresponds to the elbow of the plot. **C.** Histogram of sample correlation (all x all) in RNAseq data using two RNAseq duplicates (#13 and #6) created with the scale factor $\alpha=1$ based on the simulation from B.
### Supplementary Table 2. Characterization of methods used in challenge submissions

#### Sub-challenge 1

| Task                  | Model          |
|-----------------------|----------------|
| protein-clinical attribution matching | Logistic regression |
|                       | Random forest  |
|                       | SVM            |
|                       | Ensemble based |
|                       | KNN            |
|                       | PLS            |

#### Sub-challenge 2

| Task                  | Model          | RNA-protein matching          |
|-----------------------|----------------|-------------------------------|
| protein, RNA-clinical attribute matching | Logistic regression | regression |
|                       | Linear regression | correlation analysis |
|                       | Random forest    | Random Forest                 |
|                       | SVM              | No                            |
|                       | Ensemble based   |                               |
|                       | KNN              |                               |
|                       | PLS              |                               |

SVM: Support Vector Machine; KNN: K-Nearest Neighbors; PLS: Partial Least Squares

Red: methods used by top performers
**Supplementary Table 3.** Summary of data types and platforms of 6 real-case datasets

| Dataset          | Organism | Sample types   | # of samples | Datatype (measurement) | # of errors identified | Type of errors identified |
|------------------|----------|----------------|--------------|-------------------------|------------------------|--------------------------|
| preQC CPTAC LUAD | Human    | Tumor tissue   | 107          | RNAseq (RPKM) TMT proteomics (intensity ratio) CNV (CBS) | 8                      | 4 swapping pairs in proteomics |
| preQC CPTAC CCRCC| Human    | Tumor tissue   | 77           | RNAseq (FPKM) TMT proteomics (MS intensity) CNV (CBS) | 3                      | 3 shifts in proteomics |
| TCGA BRCA        | Human    | Tumor tissue   | 521          | RNAseq (RSEM) microarray (log2(intensity)) CNV (CBS) | 16                     | 8 swapping pairs in microarray |
| CCLE             | Human    | Cancer cell lines | 371         | RNAseq (RSEM) TMT proteomics (intensity ratio) CNV (CBS) | 0                      | N/A |
| Battle et al.    | Human    | Cell lines     | 62           | RNAseq (CPM) SILAC proteomics (intensity ratio) Riboseq (ratio) | 3                      | 1 swapping pair in RNAseq, 1 duplication in proteomics |
| Chick et al.     | Mouse    | Normal liver tissues | 192        | RNAseq (rankz transformed abundance), TMT proteomics (intensity) | 20                     | 9 swapping pairs, 2 duplication in proteomics |
### Supplementary Table 4. Gender information of the 18 samples swapped between RNAseq and protein profiles in Chick et al.

| TMT Batch | Sample name | Annotation (given) | RNAseq (predicted) | Proteomic (predicted) | Type of Swapped data |
|-----------|-------------|--------------------|--------------------|-----------------------|-----------------------|
| S14       | s_130FS    | F                  | 0                  | F                     | 0.0039                |
| S15       | s_140FH    | F                  | 0                  | F                     | 0.0147                |
| S14       | s_131FS    | F                  | 0                  | F                     | 0.0072                |
| S15       | s_141FH    | F                  | 0                  | F                     | 0.0089                |
| S14       | s_132FS    | F                  | 0                  | F                     | 0.0043                |
| S15       | s_142FH    | F                  | 0                  | F                     | 0.0129                |
| S14       | s_133FH    | F                  | 0                  | F                     | 0.0163                |
| S15       | s_143FH    | F                  | 0                  | F                     | 0.0103                |
| S14       | s_135MS    | M                  | 1                  | M                     | 0.8622                |
| S15       | s_146FS    | F                  | 0                  | F                     | 0.0112                |
| S14       | s_136MS    | M                  | 1                  | M                     | 0.9955                |
| S15       | s_147FH    | F                  | 0                  | F                     | 0.0565                |
| S14       | s_137MS    | M                  | 1                  | M                     | 0.9989                |
| S15       | s_148MH    | M                  | 1                  | M                     | 0.9948                |
| S14       | s_138MH    | M                  | 1                  | M                     | 0.9948                |
| S15       | s_149FH    | F                  | 0                  | F                     | 0.0108                |
| S14       | s_139MS    | M                  | 1                  | M                     | 0.9944                |
| S15       | s_150FH    | F                  | 0                  | F                     | 0.0167                |

Notes: TMT_14 samples are from TMT batch S14, TMT_15 samples are from TMT batch S15.
Supplementary Table 5. Gender and immune subtypes of 4 swapping pairs in preQC CPTAC LUAD dataset

| 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     |
Supplemental Experimental Procedures

I. Description of methods of top performing teams in sub-challenge 2

During the challenge event, every participant teams were requested to submit a brief description of their methods along with the submission of their challenge results. The description from the three winning teams from Lund University, Wright State University and Sentieon Inc are included below.

Team 1: Lund University

Data preprocessing

No dimensionality reduction appeared to be needed and applying different imputation techniques did not improve the results.

Feature selection

Prior to applying machine learning to predict gender and MSI, feature selections are beneficial both for identifying and correcting sample swaps, shifts and duplications and also for training the gender and MSI prediction models. In the first case, the training set was used to identify a feature set that displayed high correlation (Spearman) between RNA and proteins. This feature set was then employed in the test set. Secondly, in order to classify gender, only genes and gene products located on the X and Y chromosomes were considered. For MSI status prediction, the known genes for this type of disease were used to identify a set of Gene Ontology (GO) terms known to have association with microsatellite instability in colorectal cancer. All genes and proteins from the datasets present in any of the enriched GO terms were included as features in the respective models.

Sample matching method

Correct RNA/protein sample pairs were identified using the diagonal in the correlation matrix. Samples were assumed correctly paired when the RNA profile for data entry A, R(A), correlated more with the protein profile P(A) than any other protein profile. The correlation matrix was then updated by identifying a new subset of RNA/Protein analytes with highest correlation in the assumed correctly paired test set samples. Candidates for sample swapping errors were identified using a correlation analysis strategy similar to how correct RNA/Protein pairs were initially identified.

Sample shifting correction was performed using a strategy that depended on being seeded with a duplication event, assuming that P(B) is shifted if a duplicate of P(A) is found in its place. The location to where P(B) is shifted was identified using correlation analysis. However, since we expected a shifting chain to be four samples long, we did not assume that each of the shifts could be identified by using only the top correlating sample. Instead, we developed an algorithm that identified the top five correlating samples of each of them for each step in the chain. The search was performed according to the following logic: 1) If R(B) is a duplicate of R(A), assume R(B) occurs elsewhere (RNA sample B is shifted). 2) Find top five RNA profiles correlating with P(B): R(B1, B2, B3, B4, B5). 3) For each P(Bn) find five candidate RNA profiles R(Cn1…Cn5) by correlation. 4) For each P(Cnm), find five candidate RNA profiles R(Dnm1…Dnm5) by correlation. The search described above results in 125 (5*5*5) different four samples long candidate shift-chains (A → B → C → D). Each step in each chain was taken due to a protein and RNA sample having a correlation rank of at least five among all samples in the test set. An ideal protein sample shift chain could thus be A → B1,1 → C1,1 → D1,1, where the indices indicate the rank of RNA to protein and protein to RNA correlation, respectively. As the search depth in each step was set to five, the weakest chain that could theoretically be found would be A → B5,5 → C5,5 → D5,5. Summing the ranks yields an index ranging from 6 to 30. The 125 chains were subsequently filtered by excluding chains that did not span four patients (e.g. where the protein sample labeled B was found when searching the RNA sample labeled B, which is the case for most correct sample pairs). In the scope of this challenge, prior knowledge included approximate length of the shift/chain and that samples already swapped could not also be shifted. Hence, solutions that included already swapped or duplicated samples could also be filtered out. The remaining chains were sorted by the rank sum of all correlation steps taken, producing the most likely shift chain (having the lowest rank sum). The highest scoring protein shift chain was found as A → B1,1 → C1,1 → D1,1, and the highest scoring RNA shift chain as A → B2,2 → C1,1 → D1,1. Each top scoring chain scored significantly better than the second best. The classifier error sum for the
original clinical data notation and after correcting according to the suggested shift was then compared as validation, showing that both chains dramatically improved the classification errors.

**Team 2: Wright State University**

**Data preprocessing**

Missing values of both RNA-seq and proteomics data were handled differently. For sex-linked genes, the missing values were 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 with missing values in RNA-seq data were removed, resulting in a removal of 27% of genes; while missing values in proteomics data were imputed via Random Forest missing data algorithm.

**Feature selection and sample matching method**

1) Clinical Attributes Prediction

Four logistic regression (LR) models were trained to predict clinical attributes, gender and microsatellite instability (MSI) status, from RNA-seq and proteomics data (Table SI1). There are two major issues in model training: class imbalance and high dimensionality. The weightage was used in model training to alleviate the class imbalance issue. A higher weighting is given to the class with fewer training instances so that each class contributes equally to the loss function. As for high dimensionality issue, regularization methods such as L1 (Least Absolute Shrinkage and Selection Operator, LASSO\(^2\)) regression and elastic net were implemented to reduce the number of features.

| Input data | Transcriptomic | Proteomic |
|------------|----------------|-----------|
| Features   | Sex-linked genes | Autosomal genes | Sex-linked proteins | Autosomal proteins |
| Regularization Method | L1 | Elastic net | L1 | Elastic net |
| Output     | Gender | MSI status | Gender | MSI status |
| 10-fold CV accuracy | 97.5% | 100% | 98.75% | 100% |

Table SI1: The design and training of each classifier model

2) RNA-seq and Proteomics Samples Matching

Sample matching was performed to determine the corresponding pair of RNA-seq and proteomics sample. We used correlation as the distance metric and the expression values of highly correlated genes between RNA-seq and proteomics data were treated as a vector for every sample. Every RNA-seq sample should have the highest correlation with its corresponding proteomics sample. To reduce the noise and amplify the signal in correlation matrix, a probability matrix was derived, and the corresponding pair of samples will have the highest probability to each other. The probability matrix was 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 generated 80 pairs of RNA-seq and proteomic sample which will be used to perform label correction.

3) RNA-seq and Proteomics Data Mislabeling Correction

The stable matching algorithm generated 80 pairs of samples with a matching score. The matching score is the sum of preferential rank of RNA-seq and of proteomic samples. A matching pair should have a matching score of 2 and both the samples should belong to the same patient (i.e., have the same label). On the other hand, pairing with a matching score > 2 or pairing of samples with different labels is a mismatched pair. A mismatched pair indicates RNA-seq or proteomics data mislabeling and were further inspected for label correction. The automated label correction mechanism using predicted attributes, stable matches and matching score are explained in the following subsections for different mislabeling errors.

- **Swapping**: Swapping errors are identified when two RNA-seq data are paired with the opposite proteomics data. To determine which omics data get swapped, the predicted attributes and their classification probabilities were inspected. Two error rates were computed: one represents the
difference between annotated and predicted attributes of RNA-seq data, and another one of proteomics data. The omics data with a higher error rate has their labels corrected.

- **Duplication:** Stable matching algorithm matches one RNA-seq sample to exactly one proteomic sample. Thus, the duplicated sample pair will not pair with its corresponding sample but spuriously paired with another incorrect sample. The matching score is utilized to identify the spurious pairings, usually has a matching score > 2. For every spurious pairing, the RNA-seq and proteomic data were independently checked for their highest correlation. The omics data with the highest correlation is determined to be the duplicated data and its label was corrected to the label of its highest correlated sample. If it is a duplicate of another matching sample pair, then this is a duplication error; however, if the data is a duplicate of another mismatched pair, then it is identified as shifting error and a different correction mechanism is used to correct the label.

- **Shifting:** Shifting errors were identified through the initiation of a duplication error. A shifting chain describes multiple switching of samples in a consecutive manner. A shifting chain starts with a duplicated sample, which is identified as the previous subsection. The chain is identified by iteratively inspecting the next paired sample until the chain reaches another spurious pairing. Once the shifting chain is identified, the predicted attributes of both omics data and their classification probabilities were inspected to determine which omics data get switched. Two error rates were computed: one represents the difference between annotated and predicted attributes of RNA-seq data, and another one of proteomics data. The omics data with a higher error rate has their labels corrected.

4) Clinical Data Mislabeling Correction

Samples matching identified RNA-seq and proteomics data mislabeling, but not clinical data mislabeling. To identify clinical data swapping cases, only the matching pairs were inspected. For every matching pair, two error rates were computed: one represents the difference between annotated and predicted attributes of RNA-seq data, and another one of proteomic data. The error rates were averaged and an error rate > 0.5 was suspected to be a clinical swapping error. To perform the label correction, stable matching algorithm was performed on these suspected swapping cases. The stable matching algorithm generated pairings with consistent annotated attributes and predicted attributes. The stable matching output was then used to correct the labels of clinical data.

**Team 3: Sentience Inc.**

**Data preprocessing**

In the exploratory data analysis (EDA) step of sub-challenge 1, we observed more than 30% genes having a missing data rate above 20%. To handle the missing data in the dataset, we developed two imputation strategies for the two missing data types: Missing At Random (MAR), and Missing Not At Random (MNAR) respectively. An example of MNAR in our case is the large amount of missing protein measurements of the Y-linked genes for female samples. On the other hand, the result of MAR data might be from the system noise in the mass spectrometry mapping process. The strategy to address MNAR data is to fill the missing entries with 0, assuming that the protein abundance is either 0 or too small to be detected. The strategy for MAR data is to fill the missing data on autosomal genes by Non-negative Matrix Factorization (NMF), with the assumption that the MAR data could be partially inferred by existing data. In sub-challenge 2, however, no data imputation was applied, since with more RNA data available, we believe that replacing missing data by 0 is enough for us to detect the signals.

**Feature selection**

For this high-dimension low-sample-size dataset, we implemented both univariate feature screening and regularized model fitting to reduce the number of feature dimensions. One-way Analysis of Variance (ANOVA) was carried out in the univariate screening process, where the statistical significance of gene abundance between different clinical-labeled groups was tested. The genes with low p-values in the F-test were selected as features. Both Bonferroni correction and Benjamini-Hochberg process were performed to adjust the thresholds of p-values in the multiple testing scenarios, resulting in different feature selections for model building and stacking in the next step. Regularized models used in our approach includes the L1 regularized classifiers such as LASSO and Nearest Shrunken Centroid (NSC), enabling automatic feature selection in the fitting process. Hyper-parameters for the predictive models are chosen
through cross-validation. Figure SI1 shows the genes selected by NSC to predict MSI status using protein abundance.

![Figure SI1: Proteomics measurements selected by NSC and the class centroid for MSI status](image)

**Sample matching method**

1) Predictive models for clinical labels

LASSO, NSC and $k$-Nearest Neighbors ($k$-NN) were used to predict clinical labels. Apart from those, due to the infrequent nature of mislabeling events, we proposed a label-weighted $k$-Nearest Neighbors ($k$-NN) classifier to incorporate more data as training, giving different weights to data with different confidence levels. To build the clinical label classifiers, two schemes could be chosen: 1) Two separate binary classifiers for gender and MSI status, with predicted probabilities combined together to get a multi-class label mismatch classifier, and 2) One multi-class classifier for 4 clinical label classes: Female MSI-Low, Female MSI-High, Male MSI-Low, Male MSI-High.

As a final step, different combinations of missing data imputation strategies, feature selection methods, prediction algorithms were added to the ensemble. Since there was a pronounced shift in labels’ prior probabilities from the training data to testing data, we adjusted the posteriori probabilities to reflect the Female/Male and MSI-High/MSI-Low ratio changes, following what Saerens et al. suggested.

2) Distance Matrix for Protein-RNA relation

To further correct the mislabeled samples, we jointly analyzed measurements from both mass spectrometry and RNA-Seq by mapping two data types to each other. For $N$ samples, we defined a $2N \times N$ distance matrix $D$, with the $(i, j)$ entry representing the distance between protein sample $i$ and RNA sample $j$. Regression models were firstly built for each gene using the training set. R-squared was used as the Goodness-of-fit measurement to select a set of genes that we believe have a tighter bond in protein and RNA abundance. With the chosen gene set $G$, two novel definitions of the distance between proteomic and transcriptomic profiling data were proposed:

- **Linear Model based distance**: for each gene $k$, a linear model was built to predict protein abundance based on RNA-seq measurements. The Protein-RNA distance defined for each gene was the difference between predicted protein abundance and the observed protein abundance. The $(i, j)$ entry of the distance matrix was defined as the root of the sum of squared linear-model-based-distance over all genes.
• Rank based distance: For each gene, we defined protein rank as the ranking among protein abundance of all samples for that gene. Similarly, the RNA rank was defined as the ranking among RNA abundance of that gene among all samples. The distance of Protein and RNA on each gene was then defined as the rank difference for the compared samples. As a result, the \((i, j)\) entry of the distance matrix was then defined as the root of the sum of squared rank-based-distance over all genes.

In order to find the matched protein and RNA profiles, we searched for the shortest distance in the distance matrix. For example, the matching RNA profile of protein sample \(i\) will be RNA sample \(\text{arg min}_j (D_{ij})\). Similarly, the matching Protein profile of RNA sample \(j\) will be the sample \(\text{arg min}_i (D_{ij})\) in the proteomics data set. Inspired by the idea of Consensus Clustering\(^6\), we randomly sampled the genes used to build the Distance Matrix. We performed multiple iterations of sub-sampling, with the aim of reducing the random noise from the data set and obtaining a more robust Protein-RNA matcher. As a validation, both the linear model based distance matrix and the rank based distance matrix were able to perfectly correct the mismatch of protein and RNA profiles in training data.
II. COSMO performances on simulated datasets with different numbers of samples and features

First, to access the utility of COSMO on datasets with smaller sample sizes, simulations were performed with respect to various sample sizes \((n = 20, 30, 50, 60, 80, 100)\). For each sample size, 20 simulated datasets were generated from the CPTAC Clear Cell Renal Cell Carcinoma datasets. For each simulated dataset, 20% of samples were randomly selected to introduce mislabeling errors. The type of errors was random across different datasets to mimic more real case scenarios. COSMO was applied to these datasets, and its performance was evaluated using \(F_1\) scores. COSMO achieved a median \(F_1\) score of 1.0 in all categories across different sample sizes (see Figure SI2 and Table SI2 below), demonstrating its applicability for small sample sized datasets. Of note, COSMO achieved a lower mean \(F_1\) score in average level when sample size \(n = 20\) (\(F_1\) score = 0.856) than when sample size is larger, \(n \geq 30\) (\(F_1\) score > 0.972).

![F1 score vs Sample Size](image)

**Figure SI2.** Performance of COSMO on simulated datasets with different numbers of samples from 20 to 100. Boxes indicate the \(F_1\) score distributions among 20 random sets for different sample sizes.

| Sample Size | Sample level | Data level | Correction level | Average |
|-------------|--------------|------------|------------------|---------|
|             | Mean | Median | Mean | Median | Mean | Median | Mean | Median |
| 20          | 0.889 | 1.000 | 0.844 | 1.000 | 0.835 | 1.000 | 0.856 | 1.000 |
| 30          | 0.995 | 1.000 | 0.985 | 1.000 | 0.960 | 1.000 | 0.980 | 1.000 |
| 50          | 1.000 | 1.000 | 1.000 | 1.000 | 0.929 | 1.000 | 0.976 | 1.000 |
| 60          | 0.998 | 1.000 | 0.983 | 1.000 | 0.933 | 1.000 | 0.972 | 1.000 |
Table SI2. Mean and median of F1 score among 20 random simulated sets for each sample size.

| Sample size | Mean F1 score | Median F1 score | Mean F1 score | Median F1 score | Mean F1 score | Median F1 score | Mean F1 score | Median F1 score |
|-------------|---------------|----------------|---------------|----------------|---------------|----------------|---------------|----------------|
| 80          | 0.997         | 1.000          | 0.990         | 1.000          | 0.936         | 1.000          | 0.974         | 1.000          |
| 100         | 0.990         | 1.000          | 0.990         | 1.000          | 0.923         | 1.000          | 0.968         | 1.000          |

Then, we assessed the performance of COSMO on simulated datasets with different numbers of features (correlation > 0.5). For each feature size ($f = 20, 50, 80, 100, 150, 200, 500, 1000$), 20 datasets with 100 samples were randomly generated. COSMO was applied on these datasets and performance of COSMO was evaluated with F1 score (Figure SI3 and Table S13). Figure SI3 below shows the mean F1 score in correction level increased drastically when the subset size increased from 20 to 50 (from 0.344 to 0.781). High accuracy (median F1 score > 0.9) was achieved for all categories when the number of features were more than the number of samples ($n>=150$).

Figure SI3. Performance of COSMO on simulated datasets with different number of samples from 20 to 100. Boxes indicate the F1 score distributions among 20 random sets for different sample sizes.

| Number of correlated genes | Sample level | Data level | Correction level | Average |
|----------------------------|--------------|------------|-------------------|---------|
|                            | Mean | Median | Mean | Median | Mean | Median | Mean | Median |
| 20                         | 0.584 | 0.580 | 0.491 | 0.481 | 0.345 | 0.344 | 0.473 | 0.464 |
| 50                         | 0.879 | 0.870 | 0.855 | 0.870 | 0.764 | 0.781 | 0.833 | 0.844 |
| 80                         | 0.939 | 0.952 | 0.931 | 0.952 | 0.860 | 0.860 | 0.910 | 0.906 |
Table SI3. Performance of COSMO on simulated datasets with different numbers of correlated genes.

| Number | 100 | 150 | 200 | 500 | 1000 |
|--------|-----|-----|-----|-----|------|
|        | 0.975 | 0.982 | 0.986 | 0.987 | 0.990 |
|        | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
|        | 0.968 | 0.979 | 0.978 | 0.987 | 0.990 |
|        | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
|        | 0.873 | 0.898 | 0.914 | 0.920 | 0.922 |
|        | 0.883 | 0.906 | 0.961 | 0.976 | 0.974 |
|        | 0.939 | 0.953 | 0.959 | 0.965 | 0.967 |
|        | 0.952 | 0.958 | 0.978 | 0.976 | 0.983 |
III. Results of COSMO application on real life datasets

We applied COSMO’s performance in six real multi-omics studies (Supplementary Table 3): three datasets in which mislabeling samples were observed previously before and three other datasets without previous knowledge of existing mislabeling errors.

Dataset 1: preQC CPTAC LUAD

CPTAC initiated the effort of studying lung adenocarcinoma (LUAD) and collected multiple types of omics data and imaging data of lung tissues from LUAD patients\(^7\). Upon the collection of data, the omics data were manually inspected by checking sex annotation and performing sample alignment. Manual inspection identified four pairs of swapping of proteomic samples and the errors were corrected upon publishing. We applied COSMO’s performance in preQC LUAD dataset (the original dataset before performing quality control, QC, and still retaining the errors identified from manual inspection).

The dataset consists of RNAseq, proteomics and CNV profiles of the same 107 lung tumor samples. We performed three pairwise alignments of different omics data, each alignment involves two different types of omics data. No mislabeling found when aligning RNAseq data with CNV data (Figure SI4B), indicating these two omics data are labeled correctly. However, when proteomic data was aligned against RNAseq (Figure SI4A) and against CNV data (Figure SI4C), same four pairs of swapping were identified. This suggested all four pairs of swapping occurred in proteomic samples (Table SI4). COSMO was able to reproduce the findings which are consistent with the findings of manual inspection. Besides, COSMO confirmed that there were no additional labeling errors in the published CPTAC LUAD dataset.

| Pairwise alignment | RNAseq vs Proteomic | RNAseq vs CNV | Proteomic vs CNV |
|-------------------|---------------------|---------------|-----------------|
| Number of samples | 107                 | 107           | 107             |
| Number of features| 7443                | 18567         | 7550            |
| Number of selected features (cor > 0.5) | 4517 | 5570 | 971 |
| Number of mislabeled samples | 8 | 0 | 8 |

Figure SI4: Multiple pairwise alignments identified the swapping errors of proteomic samples.
| Type of Error       | 4 pairs of swapping | NA       | 4 pairs of swapping |
|--------------------|---------------------|----------|---------------------|
| Mislabeled Samples | Four pairs of proteomic swapping: |          |                     |
|                    | - Swapping of C3N.00559 and C3N.02572 |          |                     |
|                    | - Swapping of C3L.00510 and C3L.00093 |          |                     |
|                    | - Swapping of C3L.01330 and C3N.00203 |          |                     |
|                    | - Swapping of C3L.02219 and C3N.02586 |          |                     |

Table SI4: Summary of pairwise alignments on preQC CPTAC LUAD dataset.

**Dataset 2: preQC CPTAC CCRCC**

Clear cell renal cell carcinoma (CCRCC) is another cancer studied by CPTAC in large-scale effort⁸. Similar with previous dataset of LUAD, the omics data collected were manually inspected before the data analysis. Manual inspection identified a shifting chain of three samples and the errors were corrected upon publishing. We applied COSMO’s performance in preQC CCRCC dataset (the original dataset before performing quality control, QC, and still retaining the errors identified from manual inspection).

The dataset consists of RNAseq, proteomic and CNV data of 77 renal tumor samples. Three pairwise alignments were performed on different omics data. No mislabeling found when aligning RNAseq data against CNV data (Figure SI5B), indicating these two omics data are labeled correctly. However, when proteomic data was aligned against RNAseq (Figure SI5A) and CNV data (Figure SI5C), same three samples were found to be misaligned. This suggested the shifting chain is the continuous switching of proteomics samples (Table SI5). COSMO was able to reproduce the findings which are consistent with the findings of manual inspection and confirmed that there were no additional labeling errors in the published CPTAC CRCC dataset.

![Figure SI5](image-url)

Figure SI5: Multiple pairwise alignments identified the shifting errors of proteomic samples.

| Pairwise alignment   | RNAseq vs Proteomic | RNAseq vs CNV | Proteomic vs CNV |
|----------------------|----------------------|---------------|------------------|
| Number of samples    | 77                   | 77            | 77               |
| Number of features   | 5033                 | 18950         | 5035             |
Dataset 3: TCGA BRCA

In a previous study, 16 samples of microarray data in TCGA breast cancer (BRCA) dataset were identified to be swapped. We collected RNAseq, microarray and CNV data of 521 breast tumor samples from TCGA and then applied COSMO on the dataset. Three pairwise alignments were performed on different omics data. No mislabeling found when aligning RNAseq data against CNV data (Figure SI6B), indicating these two omics data are labeled correctly. However, when microarray data was aligned against RNAseq (Figure SI6B) and against CNV data (Figure SI6C), same eight pairs of swapping were identified. This suggested the swapping occurred in microarray data (Table SI6). COSMO was able to reproduce the findings which are consistent with the findings of the previous study.
### Table SI6: Summary of pairwise alignments on TCGA BRCA dataset.

| Number of samples | 521 | 521 | 521 |
|-------------------|-----|-----|-----|
| Number of features | 15655 | 18333 | 14807 |
| Number of selected features (cor > 0.5) | 12541 | 4651 | 3556 |
| Number of mislabeled samples | 16 | 0 | 16 |
| Type of Error | 8 pairs of swapping | NA | 8 pairs of swapping |

| Mislabeled Samples | Eight pairs of microarray swapping: |
|--------------------|-----------------------------------|
|                    | - Swapping of TCGA.BH.A0BA and TCGA.BH.A0DS |
|                    | - Swapping of TCGA.BH.A18K and TCGA.BH.A18T |
|                    | - Swapping of TCGA.BH.A0BS and TCGA.BH.A0BT |
|                    | - Swapping of TCGA.AR.A1AW and TCGA.AR.A1AV |
|                    | - Swapping of TCGA.BH.A0H3 and TCGA.BH.A0H |
|                    | - Swapping of TCGA.E2.A1B5 and TCGA.E2.A1B6 |
|                    | - Swapping of TCGA.AR.A1AN and TCGA.AR.A1AL |
|                    | - Swapping of TCGA.BH.A0EI and TCGA.A1.A0SD |

**Dataset 4: CCLE**

Cancer Cell Line Encyclopedia (CCLE) is the most comprehensive multi-omics data of various cancer cells. We collected the omics data of 371 samples. These are the samples which have all three types of omics data: RNAseq, proteomics, and CNV. To investigate if there are any potential labeling errors in CCLE datasets, we applied COSMO in a similar way as previous datasets. Three pairwise alignments were performed on different omics data and every sample is aligned consistently to each other in every alignment (Figure SI7). No error was found in CCLE dataset (Table SI7).
Figure SI7: Multiple pairwise alignments and no mislabeling was found in CCLE dataset.

| Pairwise alignment          | RNAseq vs Proteomic | RNAseq vs CNV | Proteomic vs CNV |
|-----------------------------|---------------------|---------------|------------------|
| Number of samples           | 371                 | 371           | 371              |
| Number of features          | 4699                | 10756         | 4816             |
| Number of selected features (cor > 0.5) | 1855                | 3269          | 1954*            |
| Number of mislabeled samples | 0                   | 0             | 0                |

* features with correlation > 0.25 are selected due to low feature correlation between proteomic and CNV data

Table SI7: Summary of pairwise alignments on CCLE dataset.

**Dataset 5: Battle et al.**

Battle et al. (2015) investigated the regulatory relationship of genetic variants with mRNA expression (eQTLs), ribosome occupancy (rQTLs) and protein abundance (pQTLs) \(^\text{10}\). The authors collected lymphoblastoid cell lines (LCLs) derived from humans and characterized three types of omics data: RNAseq, riboseq and proteomic data. Three pairwise alignments were performed on different omics data. In overall, three samples were found to have mislabeling error.

Two samples (GM19192 & GM19128) are found to be aligned correctly when aligning proteomic data against riboseq data (Figure SI8C). However, the RNAseq data was aligned against proteomic (Figure SI8A) and against riboseq data (Figure SI8B), these two samples found to be swapped with each other. Swapping of the RNAseq data of these two samples was identified (Table SI8).

Another duplication error was found on Sample GM19143. When RNAseq data was aligned with riboseq data (Figure SI8B), the sample was aligned perfect to itself. However, when the proteomic data
was aligned with RNAseq (Figure SI8A) and against riboseq data (Figure SI8C). Sample GM19143 is highly correlated to Sample GM19102 in both alignments, suggesting the proteomic Sample GM19143 is actually a duplicate of proteomic Sample GM19102.

Figure SI8: Multiple pairwise alignments identified one swapping pair of RNAseq sample and one duplication error of proteomic sample.

| Pairwise alignment | RNAseq vs Proteomic | RNAseq vs Riboseq | Proteomic vs Riboseq |
|--------------------|---------------------|-------------------|---------------------|
| Number of samples  | 62                  | 71                | 60                  |
| Number of features | 4340                | 13239             | 4332                |
| Number of selected features (cor > 0.5) | 257       | 1450              | 327                 |
| Number of mislabeled samples | 3           | 2                 | 1                   |
| Type of Error      | 1 pair of swapping 1 duplication | 1 pair of swapping | 1 duplication |
| Mislabeled Samples | One swapping pair of RNAseq samples (GM19192 and GM19128) One duplication error of proteomic sample: - Sample GM19143 is duplicate of Sample GM19102 |

Table SI8: Summary of pairwise alignments on dataset collected from Battle et al. (2015).
Dataset 6: Chick et al.

Lastly, we applied COSMO on a proteogenomic dataset collected from liver tissue of mice. Chick et al. (2016) studied the relationship of genetic variant with transcriptome and proteome. The liver tissues were collected from 192 individual mouse and were characterized for two omics data: RNAsseq and proteomics. We collected the dataset and applied COSMO on it to investigate if there is any potential mislabeling. COSMO identified a total 20 mislabeled samples, consisting 9 pairs of swapping and 2 duplication errors (Figure SI9A-B and Table SI9).

The sex labels were predicted based on RNAsseq and Proteomics data. The predicted sex label and alignment matching were utilized to identify the source of errors. Both duplication errors were found to be proteomics duplicate (Figure SI9C). Proteomic data of Samples s_145FH and s_134MH were actually duplicates of Samples s_117MH and s_65MS, respectively.

As for the nine swapping pairs, four among them occurred between opposite sex and COSMO identified these four swapping occurred in proteomics data (Supplementary Table 4). The other five swapping pairs occurred between same sex and COSMO could not identify the source of error. We further investigated these 20 samples and found out that all the mislabeled samples were from two multiplexed tandem mass tag (TMT) experiments, TMT batch S14 and S15. All nine swapping involve two samples of different TMT batch. Thus, it is hypothesized that the entire TMT batch get swapped during proteomics characterization. The errors were confirmed by the authors of the publication.

Figure SI9: COSMO identified 20 mislabeled samples, consisting of 9 pairs of swapping and 2 duplication errors.

| Pairwise Alignment | RNAsseq vs Proteomic |
|--------------------|----------------------|
| Number of samples  | 192                  |
| Number of features | 6361                 |
| Number of selected features (cor > 0.5) | 1324 |
| Number of mislabeled samples | 20 |
| Type of Error | 9 pairs of swapping  
|              | 2 duplication |

Table SI9: Summary of pairwise alignments on the mouse proteogenomic dataset.

Overall, COSMO was applied on 6 real life proteogenomic datasets which most of them have both RNA-seq and proteomics data with other omics data such as microarray, Riboseq and CNV. We observed that mislabeling occurred in 5 out of 6 datasets, demonstrating that mislabeuling is not uncommon across datasets. The number of mislabeled samples varied across datasets and the mislabeling rate ranges from 0 to 10.42%, with an overall rate of 3.73% (Supplementary Table 3). Our investigation showed that human errors are not uncommon and, in some cases, one error made in upstream workflow could have a devastating effect on downstream workflow, such as what happened in the mouse proteogenomic study. One swapping occurred between two different TMT batches at the beginning of proteomic mass spectrometry and this caused 9 samples get swapped with another 9 samples, leading to a high mislabeling rate observed at later stage of analysis.
IV. COSMO User's manual (Version 1.0.0)

1 Introduction
COSMO (https://github.com/bzhanglab/COSMO) is a tool for sample mislabeling correction using multi-omics datasets. For specific bug reports or issues please use the Github issues tracker (https://github.com/bzhanglab/COSMO/issues).

2 Required resources
OS: Linux or Mac OS.
Hardware: 4 CPUs (more is better), 4 Gb memory.
Input data for testing: A testing dataset which contains both proteomics and RNA-Seq data is available at https://github.com/bzhanglab/COSMO/tree/master/example_data.
Installation:
(1). Download COSMO; git clone https://github.com/bzhanglab/cosmo
(2). Install Docker (>=19.03) following the instruction at https://docs.docker.com/engine/install/;
(3). Install Nextflow (>=20.10.0) following the instruction at https://www.nextflow.io/.

All other tools used by COSMO have been dockerized and will be automatically installed when COSMO is run in the first time on a computer. On Mac OS, users may need to increase the resource limitation for both CPU and memory to run Docker by following the instruction at https://docs.docker.com/docker-for-mac/.

3 Usage
3.1 Command line options
$ nextflow run cosmo.nf --help

$ nextflow run cosmo.nf

The number of CPUs used for the analysis can be set by changing the value in the nextflow config file as shown below. In default the limitation of CPUs number is 8.

```yaml
docker {
  enabled = true
  remove = true
  numOptions = "-u $(id -u):$(id -g)" 
  temp = "/tmp/"
}
process {
  cpus=8
}
```

3.2 Input file format
To run COSMO, two different types of omics datasets must be provided. Currently we support proteomics, transcriptomics or CNV data. The formats for both types of datasets (--d1_file, --d2_file) are the same and should be provided at gene level in tsv format (tab-delimited file). An example input of quantification dataset (--d1_file or --d2_file) is shown below. The first column is the gene ID and all the other columns are the expression of proteins at gene level in different samples. Missing value should be present as "NA". The
quantification data should be well normalized before running COSMO. Missing value imputation is not required to run COSMO. A complete example dataset could be found at https://github.com/bzhanglab/COSMO/tree/master/example_data.

|    | Testing_1 | Testing_2 | Testing_3 | Testing_4 | Testing_5 | Testing_6 | Testing_7 | Test |
|----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|------|
| A1 BG | 1.5963    | 2.8484    | 2.1092    | 2.7922    | 2.4444    | 3.9907    | 3.6792    | 3.73 |
| A2M  | 5.9429    | 5.0089    | 6.0823    | 6.0093    | 6.4553    | 6.0097    | 6.014     | 6.97 |
| AAAS | 1.9337    | 2.951     | 3.5984    | 2.0419    | 2.1217    | 0.9662    | 1.0086    | NA  |
| AACS | 1.7549    | NA        | 2.3948    | NA        | 0.9946    | 2.5969    | NA        | NA  |
| AAGAB| NA        | NA        | 0.9982    | NA        | 1.0282    | 1.6296    | NA        | NA  |
| AAKI | 1.0459    | 2.5435    | 1.7449    | NA        | 1.0653    | 0.9855    | 2.0395    | 1.15 |

The input for parameter --sample_file is the sample annotation file in tsv format (tab-delimited file), and an example is shown below. This file is used in clinical attribute prediction. One or more attributes can be selected. A complete example dataset could be found at https://github.com/bzhanglab/COSMO/tree/master/example_data.

| sample | age | gender | stage | colon_rectum | msi | tumor_normal |
|--------|-----|--------|-------|--------------|-----|--------------|
| Testing_1 | 47  | Female | High  | Colon        | MSI-Low/MSS | Tumor        |
| Testing_2 | 68  | Female | High  | Rectum       | MSI-Low/MSS | Tumor        |
| Testing_3 | 52  | Male   | Low   | Colon        | MSI-Low/MSS | Tumor        |
| Testing_4 | 54  | Female | Low   | Colon        | MSI-High   | Tumor        |
| Testing_5 | 72  | Male   | High  | Colon        | MSI-Low/MSS | Tumor        |

3.3 Running COSMO
Below is an example to show how to run COSMO on the example data at https://github.com/bzhanglab/COSMO/tree/master/example_data:

```
nextflow run cosmo.nf --d1_file example_data/test_pro.tsv \
--d2_file example_data/test_rna.tsv \
--cli_file example_data/test_cli.tsv \
--cli_attribute "gender,msi" \
--out_dir output
```

3.4 Output of COSMO
The output folder of COSMO contains four folders. An example folder structure for the above example is shown below:

```
output/
  -- data_use
  -- final_res_folder
  -- method1_folder
  -- method2_folder
4 directories
```

A detailed description for each folder based on the output from the above example is shown below:

1. Folder **"data_use"**: this directory contains all the input data files as shown below.
Folder “method1_folder”: this directory contains files generated from method 1.

- cleaned_data1.tsv
- cleaned_data2.tsv
- clinical_attributes_pred.tsv
- error.tsv
- genes.tsv
- pairwise_matching.tsv
- sample_correlation.csv
- sample_correlation.png

Folder “method2_folder”: this directory contains files generated from method 2.

- cleaned_data1.tsv
- cleaned_data2.tsv
- clinical_attributes_pred.tsv
- error.tsv
- genes.tsv
- pairwise_matching.tsv
- sample_correlation.csv
- sample_correlation.png

File genes.tsv: Chromosomes annotation of genes.
File cleaned_data1.tsv: Preprocessed data from the first dataset, d1_file (Missing value imputed if any).
File cleaned_data2.tsv: Preprocessed data from the second dataset, d2_file (Missing value imputed if any).
File sample_correlation.csv: Pearson correlation between samples from the first dataset (Rows) and the second dataset (Columns).
File sample_correlation.png: Heatmap image file of `sample_correlation.csv`.
File pairwise_matching.tsv: Matching generated by stable marriage correlation. Every row indicates one matching pair of samples from the first dataset (d1_label) to samples from the second dataset (d1_label). The column d1rank is the preferential rank of d1 sample matched to the d2 sample; d2rank is the preferential rank of d2 sample matched to the d1 sample.

| d1   | d1_label   | d2   | d2_label   | d1rank | d2rank | distance | correlation |
|------|------------|------|------------|--------|--------|----------|-------------|
| 1    | Testing_1  | 1    | Testing_1  | 1      | 1      | 2        | 0.60889     |
| 2    | Testing_2  | 2    | Testing_2  | 1      | 1      | 2        | 0.59604     |
| 3    | Testing_3  | 3    | Testing_3  | 1      | 1      | 2        | 0.64042     |
| 4    | Testing_4  | 4    | Testing_4  | 1      | 1      | 2        | 0.76045     |
| 5    | Testing_5  | 5    | Testing_5  | 1      | 2      | 3        | 0.66900     |
| 6    | Testing_6  | 6    | Testing_7  | 1      | 1      | 2        | 0.77152     |
| 7    | Testing_7  | 7    | Testing_6  | 1      | 1      | 2        | 0.70996     |
| 8    | Testing_8  | 8    | Testing_8  | 1      | 1      | 2        | 0.69787     |
| 9    | Testing_9  | 9    | Testing_9  | 1      | 1      | 2        | 0.75336     |

File clinical_attributes_pred.tsv: Classification results of every samples for both datasets, using method of winning team 1. Column gender_prob is the annotated binary label, d1gender_prob is the predicted probability of sample from the first dataset; while d2gender_prob is of sample from the second dataset. More columns will be generated if there are more clinical attributes.
errors.tsv: count of different types of mislabeling errors.
final.tsv: table of corrected labels. Any inconsistency of id in the same row indicates the presence of mislabeling error.

| sample | gender | gender_prob | d1gender | d1gender_prob | d2gender | d2gender_prob |
|--------|--------|-------------|----------|---------------|----------|---------------|
| Testing_1 | Female | 0           | Female   | 0.01724       | Female   | 0.32446       |
| Testing_2 | Female | 0           | Female   | 0.00930       | Female   | 0.17867       |
| Testing_3 | Male   | 1           | Male     | 0.97656       | Male     | 0.78810       |
| Testing_4 | Female | 0           | Female   | 0.00489       | Female   | 0.25205       |
| Testing_5 | Male   | 1           | Male     | 0.99710       | Male     | 0.58199       |
| Testing_6 | Male   | 1           | Male     | 0.99831       | Female   | 0.41568       |
| Testing_7 | Female | 0           | Female   | 0.02782       | Male     | 0.57772       |
| Testing_8 | Male   | 1           | Male     | 0.99377       | Male     | 0.76312       |
| Testing_9 | Male   | 1           | Male     | 0.99856       | Male     | 0.69589       |

(3) Folder “method2 Folder”: this directory contains files generated from method 2.
test_ModelA_results.csv: classification results of every samples of the first dataset, using method of winning team 2.
test_ModelB_results.csv: classification results of every samples of the second dataset, using method of winning team 2.

(4) Folder “final_res_folder”: this directory contains files generated from combining method 1 and method 2.
cosmo_final_result.tsv: table of corrected labels. The table is generated using integrated classification results of both method_1 and method_2. Each sample is assigned a number as unique id. A row with the consistent id across all the columns, indicates all the data belongs to the same patient and there is no mislabeling.

| sample | Clinical | Data1 | Data2 |
|--------|----------|-------|-------|
| Testing_1 | 1        | 1     | 1     |
| Testing_2 | 2        | 2     | 2     |
| Testing_3 | 3        | 3     | 3     |
| Testing_4 | 4        | 4     | 4     |
| Testing_5 | 5        | 5     | 5     |
| Testing_6 | 6        | 6     | 7     |
| Testing_7 | 7        | 7     | 6     |
| Testing_8 | 8        | 8     | 8     |
| Testing_9 | 9        | 9     | 9     |

A row with different id indicates mislabeling error. Below is the example of swapping error in which samples Testing_6 and Testing_7 get swapped in Data2.
The same id occurred twice indicates duplication error. Below is an example of a duplicate sample Testing_8 in Data1.

| sample     | Clinical | Data1 | Data2 |
|------------|----------|-------|-------|
| Testing_6  | 6        | 6     | 7     |
| Testing_7  | 7        | 7     | 6     |

Shifting error are represented with a continuous switching of id. Table below shows an example of a shifting error, where Samples Testing_10, Testing_11 and Testing_12 get shifted consecutively in Data2.

| sample     | Clinical | Data1 | Data2 |
|------------|----------|-------|-------|
| Testing_10 | 10       | 10    | 10    |
| Testing_11 | 11       | 11    | 10    |
| Testing_12 | 12       | 12    | 11    |
| Testing_13 | 13       | 13    | 12    |
| Testing_14 | 14       | 14    | 14    |
REFERENCES:

1. Jonchere, V., Marisa, L., Greene, M., Virouleau, A., Buhard, O., Bertrand, R., Svrcek, M., Cervera, P., Goloudina, A., Guillerm, E., et al. (2018). Identification of Positively and Negatively Selected Driver Gene Mutations Associated With Colorectal Cancer With Microsatellite Instability. Cell Mol Gastroenterol Hepatol 6, 277-300. 10.1016/j.jcmgh.2018.06.002.

2. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267-288.

3. Cichocki, A. (2009). Nonnegative matrix and tensor factorizations : applications to exploratory multi-way data analysis and blind source separation (Chichester, U.K.: John Wiley).

4. Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statistical Science, 104-117.

5. Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural computation 14, 21-41.

6. Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning 52, 91-118.

7. 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 e235. 10.1016/j.cell.2020.06.013.