2021

Sample Mislabeling Detection and Correction in Bioinformatics Experimental Data

Soon Jye Kho
Wright State University

SAMPLE MISLABELING DETECTION AND CORRECTION IN BIOINFORMATICS EXPERIMENTAL DATA

A dissertation submitted to partial fulfillment of the requirements for the degree of Doctor of Philosophy

By

SOON JYE KHO
B.S. Biomedical Science, University Tunku Abdul Rahman, Malaysia, 2012
Master in Bioinformatics, University of Malaya, Malaysia, 2015

2021
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

July 22, 2021

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Soon Jye Kho ENTITLED Sample Mislabeling Detection and Correction in Bioinformatics Experimental Data BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Michael Raymer, Ph.D.
Dissertation Director

Yong Pei, Ph.D.
Director, Computer Science and
Engineering Ph.D. Program

Barry Milligan, Ph.D.
Vice Provost of Academic Affairs
Dean of the Graduate School

Committee on
Final Examination

Michael Raymer, Ph.D.

Tanvi Banerjee, Ph.D.

Travis Doom, Ph.D.

Michael Markey, Ph.D.

ABSTRACT

Kho, Soon Jye. PhD. Department of Computer Science and Engineering, Wright State University, 2021. Sample Mislabeling Detection and Correction in Bioinformatics Experimental Data.

Sample mislabeling or incorrect annotation has been a long-standing problem in biomedical research and contributes to irreproducible results and invalid conclusions. These problems are especially prevalent in multi-omics studies in which a large set of biological samples are characterized by multiple types of omics platforms at different times or different labs. While multi-omics studies have demonstrated tremendous value in understanding disease biology and improving patient outcomes, the complexity of these studies may increase opportunities for human error. Fortunately, the interrelated nature of the data collected in multi-omics studies can be exploited to facilitate the identification and, in some cases, correction of mislabeling errors. The dissertation proposed a pipeline comprising statistical and machine learning techniques to identify mislabeled samples and correct the sample labels. Expected correlations between copy number variation, gene transcript abundance, protein abundance and microRNA expression were used to identify mislabeled samples. In datasets with only two omics data, the label corrections were performed by exploiting gender-specific indicators of the mislabeled samples; whereas in datasets with more than two omics data, a network topology realignment method was proposed to perform label correction. We demonstrated the effectiveness of the pipeline in several cancer datasets by simulation experiments. The pipeline was then performed on several public multi-omics datasets and in overall, 2.71% of the samples are found to be mislabeled.

TABLE OF CONTENTS

1 INTRODUCTION 1

1.1 Multi Omics Studies 2

1.2 Sample Mislabeling 4

1.3 Consequences of Sample Mislabeling 6

2 LITERATURE REVIEW 10

2.1 Detecting Class Mislabeling 11

2.1.1 Classification Approaches 12

2.1.2 Statistical Approaches 16

2.2 Detecting Individual Mislabeling 17

3 MATERIALS AND METHODS 19

3.1 Data Collection 19

3.1.1 Colorectal Cancer Dataset (COAD) 19

3.1.2 Kidney Cancer Dataset (CCRCC) 20

3.1.3 Lung Cancer Dataset (LUAD) 21

3.2 Mislabeling Simulation 22

3.2.1 Swapping 22

3.2.2 Duplication 22

3.2.3 Shifting 24

3.3 Pearson Correlation Coefficient 24

3.4 Stable Matching Algorithm..25

3.5 Evaluation 26

4 MISLABELING DETECTION..29

4.1 Pairwise Alignment 29

4.1.1 Correlation Signal Extraction 29

4.1.2 Coexpression Signal Extraction for miRNA Data........................31

4.1.3 Stable Matching for Detection..32

4.2 Evaluation 33

4.2.1 Gene Correlation Inspection..33

4.2.2 Simulation Experiments 35

4.2.3 miRNA Coexpression Inspection 38

4.2.4 Simulation Experiments for miRNA Data 40

4.3 Discussion 44

5 MISLABELING CORRECTION IN TWO OMICS DATA 45

5.1 Attribute Prediction 45

5.2 Individual Label Correction 48

5.3 Simulation and Evaluation 51

5.4 Real Case Study – Mouse Proteogenomic Dataset 56

5.5 Discussion 60

6 MISLABELING CORRECTION IN MULTI OMICS DATA 62

6.1 Network Topology Realignment 62

6.2 Simulation and Evaluation 65

6.3 Real Case Study 67

6.3.1 TCGA Breast Cancer Dataset (BRCA) 67

6.3.2 Lymphoblastoid Cell Lines (LCLs) Dataset 70

6.3.3 Tuberculosis Patients Blood Gene Expression 72

6.4 Prevalence 75

7 CONCLUSION AND FUTURE WORK 81

REFERENCES 84

APPENDIX 93

LIST OF FIGURES

Figure 1.1 Number of “multi omics” related publications in PubMed 4

Figure 3.1 Simulation of duplicated samples 24

Figure 4.1 Imputation of sample correlation for performing pairwise alignment 30

Figure 4.2 Stable Matching algorithm pairs every instance from two different omics data 33

Figure 4.3 Distribution of Gene Pearson Correlation between different pairs of omics datasets 34

Figure 4.4 Simulation experiment for evaluating the performance of correlation signal in detecting mislabeled samples 36

Figure 4.5 Sample level F1 scores of mislabeling detection between different pairs of omics datasets 37

Figure 4.6 Distribution of Pearson correlation of miRNA-gene pairs 39

Figure 4.7 Sample level F1 scores of detecting mislabeled samples in pairwise alignment of miRNA data with three other omics data 41

Figure 4.8 Sample level F1 score achieved against different number of feature pairs 43

Figure 5.1 The workflow of predicting sex label of omics data 47

Figure 5.2 The automated correction algorithm for each type of mislabeling error 50

Figure 5.3 Simulation to evaluate the correction performance of COSMO 52

Figure 5.4 Number of mislabeled samples in each simulation across different datasets 53

Figure 5.5 F1 scores of label correction across different pairs of omics data 54

Figure 5.6 F1 scores against different error rates 56

Figure 5.7 Four pairs of swapping were determined to be proteomics swapping 59

Figure 5.8 LOD score of OMA1 protein expression with a genetic marker in Chromosome 4 increased from 24 to 31 upon rerun the pQTL analysis on corrected data 60

Figure 6.1 Network Topology Realignment for correcting the labels for datasets with more than two types of omics data 64

Figure 6.2 F1 scores of network topology realignment in multi omics data across different error rates 65

Figure 6.3 Number of DEGs before and after correction across different thresholds 69

Figure 6.4 Change of false discovery rate of genes before and after the correction. Horizontal and vertical grey lines indicate the FDR cutoff of 0.05 69

Figure 6.5 Distribution of gene-wise Pearson Correlation across different pairs of omics data 71

Figure 6.6 Three Mislabeled samples found in LCLs dataset 71

Figure 6.7 Mislabeling of 25 samples from 17 patients 73

Figure 6.8 Number of differentially expressed probes (left) and genes (right) between Week 4 and Week 26 before and after the label correction 74

Figure 6.9 Change of false discovery rate of genes before and after the correction. Horizontal and vertical grey lines indicate the FDR cutoff of 0.05 75

LIST OF TABLES

Table 4.1 Mean of gene Pearson correlation between two omics dataset 35

Table 4.2 Number of miRNA-gene pairs and the mean of Pearson correlation 40

Table 5.1 Mean and Median of F1 scores in sample, data and correction level 54

Table 5.2 The annotated labels and predicted labels of 18 swapped samples. 58

Table 6.1 Mean and Median F1 scores across different error rates 66

Table 6.2 Eight swapping pairs of 16 mislabeled microarray data. Each line represents each swapping pair 68

Table 6.3 Summary of mislabeled samples found in TCGA datasets 76

Table 6.4 Datasets collected from ExpressionAtlas which contain mislabeled data. Details of other datasets could be found in Appendix A 78

Table 6.5 Overall mislabeling rate compiled from various sources of datasets 80

LIST OF ABBREVIATIONS

NSCLC	Non-Small Cell Lung Cancer
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	messenger RNA
miRNA	microRNA
LC-MS	Liquid Chromatography-Mass Spectrometry
NMR	Nuclear Magnetic Resonance
TCGA	The Cancer Genome Atlas
LOO	Leave One Out
rRNA	ribosomal RNA
QTL	Quantitative Trait Loci
AUC	Area Under the Curve
SNP	Single Nucleotide Polymorphisms
NCI	National Cancer Institute
CPTAC	Clinical Proteomic Tumor Analysis Consortium
COAD	Colon adenocarcinoma
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
TMM	Trimmed mean of M-values
CCRCC	Clear Cell Renal Cell Carcinoma
LUAD	Lung Adenocarcinoma
CNV	Copy Number Variation

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CV	Cross Validation
TMT	Tandem Mass Tag
LOD	Log Odd Ratio
DEG	Differentially Expressed Gene
FDR	False Discovery Rate
CCLE	Cancer Cell Line Encyclopedia
EBI	European Bioinformatics Institute
BRCA	Breast Carcinoma
GBM	Glioblastoma Multiforme
KIRC	Kidney Renal Clear Cell Carcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma
LGG	Brain Lower Grade Glioma
LUSC	Lung Squamous Cell Carcinoma
OV	Ovarian Serous Cystadenocarcinoma

ACKNOWLEDGEMENTS

I joined Wright State University in Summer 2016 and officially joined BIRG lab in Fall 2018. First and Foremost, I would like to thank my dissertation advisor, Dr. Michael Raymer, for giving me the opportunity to join his lab and work under his guidance. I thank him for his immense support and patience throughout the period of my work here. His expertise in the research area was invaluable in formulating the research questions and methodology. He always encouraged me to think creatively and pursue any interesting idea. I am deeply grateful for the long discussions, encouragement and everything he has done for me. I also would like to thank my committee members Dr. Tanvi Banerjee, Dr. Travis Doom, and Dr. Michael Markey. They have been extremely constructive in their suggestions towards completing my thesis. Their insightful feedback pushed me to sharpen my thinking and brought my work to the next level.

I would like to thank the present and past members in the lab Hima Yalamanchili, Xiu Huan Yap, Daniel Foose, Joy Prakash Sain, Clayton Allex-Buckner, Ray Kulhanek, Aaron Hammer, Hannah Shows, Jacob Ross and Denys Guriev. I am thankful for all the discussions we had and the opportunity of learning from each other. I would also like to thank Dr. Amit Sheth and his present and past lab members Vinh Nguyen, Sanjaya Wijeratne, Lakshika Balasuriya, Sujan Perera, Sarasi Lalithsena, Amir Yazdavar, Swati

Padhee, Shreyansh Bhatt, Joey Yap, Ruwan Wickramarachchi and Thilini Wijesiriwardene and others for their help and support.

I would like to express my gratitude to the Department of Computer Science and Engineering at Wright State University for considering me as a potential student. I thank the former chair of the department, Dr Mateen Rizki, for giving me the opportunity of being a graduate teaching assistant which supported me throughout my study. I would like to thank the Computer Science and Engineering department staff, Ms. Jennifer Limoli and Ms. Wendy Chetcuti, for their cooperation and help.

Lastly, I would like to thank my family. My parents, Hock Lai Kho and Yoke Fong Pan, have always been there for me, supporting and inspiring me throughout my study. Thanks to my siblings who have cheered me up and my friends who have motivated me from time to time. Lastly, a special thanks to my dearest grandparents for their unconditional love and encouragement. This dissertation would not have been possible without their enthusiastic support of my ambitions.

1. INTRODUCTION

Omics refers to the global and comprehensive assessment of a set of biological molecules. The advances in omics technologies in the past two decades have had a profound impact on the biomedical sciences. Diverse types of omics data have been generated and studied extensively to understand complex biological processes. Transcriptomics data has been well explored in the literature for associations with specific diseases with the intent of understanding and/or predicting susceptibility to disease, morbidity, and disease progression. Ma et al. (2003) generated gene expression data of breast cancer tissues of distinct stages (premalignant, preinvasive and invasive) and distinct grades (grade I, II, and III). They found that the gene expression exhibits a similar pattern among different stages, suggesting the alteration of gene expression is already present in the preinvasive stage. In contrast with stages, a distinct gene expression pattern was observed in different grades and several genes were identified to be differentially expressed in Tumor grade III. Tonon et al. (2005) characterized the genomic profile of non-small cell lung cancer (NSCLC) and identified 319 copy number alterations (CNA). Recurrent CNAs in different samples allow the grouping into minimal common regions (MCR). The authors found 93 MCRs in total which covers a number of tumor suppressor genes and oncogenes found to be implicated in NSCLC. Rai et al. (2002) profiled peptide abundance of plasma of patients diagnosed with ovarian cancer. Four biomarkers were identified to have high discriminatory power for cancer detection.

1.1 Multi Omics Studies

Omics technology provides a high throughput approach to identify a list of differences associated with the diseases, providing insights on the different biological processes occurring in patients and normal individuals. However, the insights from single omics studies may be limited as the data only reveals the difference of one set of biological molecules and does not uncover the flow of such information to other molecules. As such, researchers may adopt a system biology approach and integrate multiple omics in studies of a variety of diseases. Over the past two decades, rapid development of omics technologies has facilitated the measurement of variations in form, abundance, and state of a wide range of biological macromolecules. Some of the more common omics technologies include:

- Genomics - assessment of DNA sequence variations among and within individuals, tissues, and cells.
- Transcriptomics - assessment of RNA transcript abundance or state. Variations exist to allow specific observation of messenger RNA (mRNA) transcripts, microRNA (miRNA) transcripts, alternative splicing of mRNA transcripts, ribosome translation of mRNA transcripts.
- Metagenomics - exploration of the collective genome of a population of organisms.
- Epigenetics - exploration of changes in genomic state that do not involve sequence variation, including chromatin structure and DNA methylation state.
- Proteomics - assessment of the relative abundance of expressed proteins.

- Phosphoproteomics - assessment of the relative abundance of proteins containing a phosphate group as a result of posttranscriptional modifications.
- Metabolomics - assessment of the presence and relative abundance of metabolites, often assayed using liquid chromatography-mass spectrometry (LC-MS) or nuclear magnetic resonance (NMR) spectroscopy.

Mertins et al. (2016) characterized genomic, transcriptomic, proteomic and phosphoproteomic profiles of 105 breast cancer samples. The authors performed integrative analysis on these omics landscapes and revealed several genomic alterations that affect the proteomic abnormality. Notably, the loss of chromosome 5q exhibits the most trans-association where loss of CETN3 and SKP1 is associated to elevated expression of epidermal growth factor receptor (EGFR), and SKP1 loss to increased SRC tyrosine kinase. On the other hand, Ding et al. (2018) conducted an ambitious study in which they inspected 11,000 tumor samples across 33 different human cancer types. The study aims to elucidate the molecular processes governing oncogenesis by uncovering the influence of somatic mutation to the carcinogenesis process. The study uncovers the association of somatic mutations with other omics (epigenome, transcriptome, and proteome) which helps in identifying the driver genes and therapeutic targets.

Multi omics studies are becoming more common in recent years. The search term “multi omics” retrieved over 1121 publications in Pubmed in 2020, compared to 17 publications in 2010 (Figure 1.1). The increasing popularity of multi omics studies has called for a repository suited for handling and disseminating this sort of system biology data. The Cancer Genome Atlas (TCGA) is a pan-cancer analysis project that has profiled and analyzed a large number of human tumors to discover molecular aberrations at the DNA, RNA, protein, and epigenetic levels (Weinstein et al., 2013). It is by far the largest such repository that has been made available for study by the scientific community.

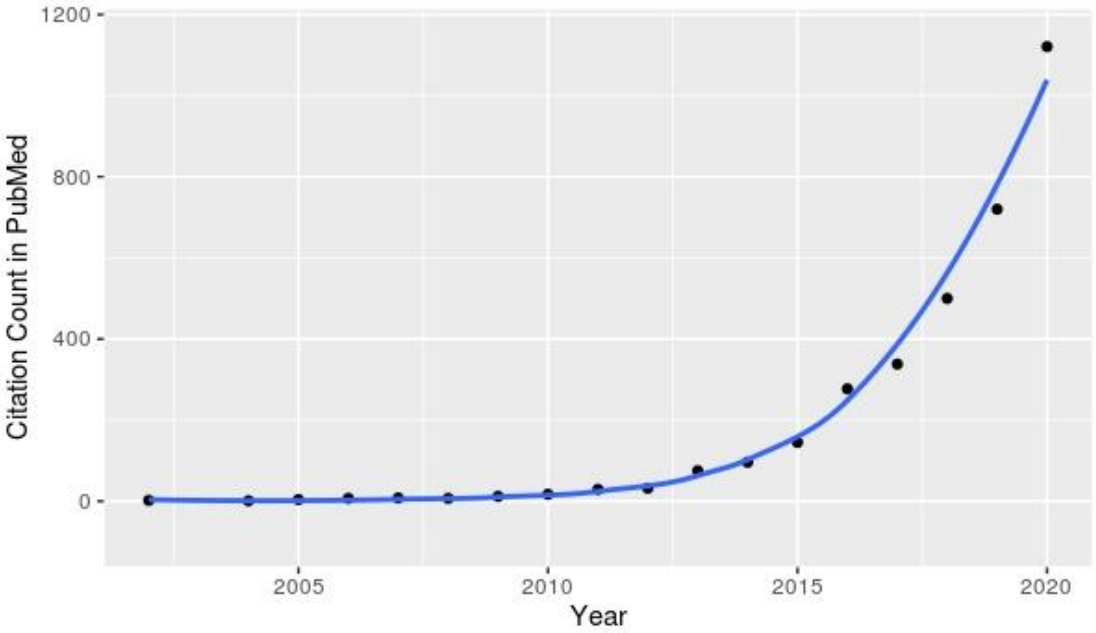


Figure 1.1: Number of “multi omics” related publications in PubMed.

1.2 Sample Mislabeling

Multi omics studies allows researchers to dissect a biological process from different aspects and understand it via a holistic approach. Consequently, a multi omics study is inevitably a large scale study that involves collaboration among researchers of different specialties.

The large scale of such study does not come without consequences. Human errors can occur at multiple steps in the experimental process including sample collection, transportation, sample analysis and data generation, and data analysis and interpretation. Sample mislabeling is particularly concerning as it might go unnoticed and can introduce significant noise into the data.

While the majority of experiments and results reported in the scientific literature are undoubtedly accurate and reproducible, there nevertheless remain numerous instances of identified experimental and procedural errors. Toker et al. (2016) looked for discrepancies between gene expression-inferred and investigator-annotated sex in 70 human microarray datasets. The authors found that among 4160 samples, 83 (2%) of them are mislabeled. These mislabeled samples are scattered across 32 datasets, showing an alarmingly high prevalence of mislabeling in datasets where 46% (32 / 70) of datasets contain mislabeled samples.

Other than biomedical research settings, sample mislabeling and data mishandling have been long-standing problems in medical settings. The U.S. Institute of Medicine (Kohn et. al, 2007) published a report entitled To Err is Human which estimates as many as 98,000 die in any given year from medical errors. The report increased the awareness of medical errors which in turn pushed the initiative for prevention and mitigation. With the intention of improving patient safety, Astion et al. (2003) investigated reported incidents in the laboratory which could potentially cause adverse events to patients. In their study, 129 incidents were reported in a 16 months period which have the potential to cause adverse events, though the errors may have been intercepted before causing harm to the patient. The authors then examined the factors for these incidents. Some factors, such as incorrect requisition, missing collection of specimens, lost or delayed specimen, suboptimal or ruined specimens, led to the failure or delay of data generation. Though unfortunate, these factors are easily recognized and corrected. In contrast, other factors such as specimen mislabeling and data entry errors are harder to detect and could cause serious harm as these factors can result in incorrect interpretation of the patient's condition. Moreover, specimen mislabeling and data entry errors are not uncommon and constitute 26% of all the incidents examined.

1.3 Consequences of Sample Mislabeling

In the medical settings, sample mislabeling may incur unnecessary patient discomfort, additional facilities and labor cost, increase morbidity and medical cost. Astion et al. (2003) reported that 5% of laboratory incidents caused actual adverse events. Valenstein et al. (2006) estimated that the rate of adverse events out of mislabeling events is ~5.29% (324 out of 6123).

In biomedical research settings, sample mislabeling may cause statistical power loss, irreproducible results, invalid conclusions and increased research cost. Statistical power loss has been a concern in the genetics research community. Simulation showed that sample mislabeling has disproportionate effects on the power to detect genetic associations in genome-wide studies, especially when the sample size is small (Buyske et al., 2009;

Edwards et al., 2005). Another study reported that sample mislabeling presents a problem in detecting genetic variants associated with diseases, specifically those variants with small genetic effect and low frequency (Samuels et al., 2009).

Correctness and reproducibility of experiments are cornerstones of the scientific method. However, these are significantly impacted by sample mislabeling. In the year 2007, Rae et al. (2007) found that the cell line MDA-MB-435, famously known as the "triple-negative breast cancer" cell line, is actually derived from a melanoma cell line. Yet, despite the identification of the error, researchers still published studies using the cell line in international peer reviewed journals over the following years. Prasad & Gopalan (2015) reported that a total of 890 published studies have used the cell line as a model for human breast cancer and 219 among them are published after the year 2007. The conclusions of notable studies that investigate the effect of drugs using this cell line may thus be questionable.

While mislabeling of one sample does not always falsify the claims of a study (depending on the total number of samples in the study), mislabeling of a significant portion of samples may invalidate the conclusion. Moloney et al. (2016) published a study which investigated the genetic underpinnings of amyotrophic lateral sclerosis (ALS) but the study was retracted later as it was discovered that the mouse line was mislabeled. Mice mislabeled as expressing wild-type MATR3 were actually expressing the mutant variant of the MATR3 gene with a mutation of F115C. While the retraction of published studies may incur unnecessary research and labor costs, this is nevertheless a preferable outcome compared to undiscovered errors resulting in false study conclusions.

Many experimental assays and protocols include quality control checks that can help to verify the correctness of the experimental conditions and instruments used. However, gross human error such as sample mislabeling or incorrect analysis assumptions remain difficult to detect and correct, especially in high-throughput multi-omics studies. Given the high prevalence and severe consequences of sample mislabeling, it is desirable to have a quality check system to ensure the correctness of data. This not only safeguards the patient safety in medical settings, but also improves reproducibility of a study and prevents any invalid claim in biomedical research settings.

Detecting individual mislabeling is a nontrivial task. However, unlike many other multi-omics problems which suffer from the well-known curse of dimensionality (Sen & Others, 2005) - decrease in classification performance as the number of independent variables increases - it may be possible to harness the plentitude of observed features typical to these studies to increase confidence for mislabeling identification and even correction. As multi-omics studies are becoming more commonplace, it is both desirable and feasible to develop a scalable automated approach.

We hypothesized that if more dimensions of data are generated, it could provide more information to determine the source of error and help in relabeling the data automatically to the individual level. Specifically, we address two research questions here: 1) Can correlation signals across different omics data accurately identify individual mislabeling errors? and 2) Is the information in typical multi-omics systems biology studies sufficient to afford automated correction of mislabeling errors?

2. LITERATURE REVIEW

Machine Learning is a method of data analysis that automates analytical model building. Machine learning approaches, as applied to data analytics, attempt to identify patterns in observations (generally referred to as instances). Each observation may comprise multiple variables (or features) collected from a single sample or individual. In cancer transcriptomics for example, a sample may consist of tens of thousands of gene expression values from a single tumor tissue sample. In supervised machine learning the learned patterns are applied to classify the observations into two or more classes (e.g. cancer tissue versus healthy tissue). Unsupervised machine learning, in contrast, does not associate a class with each sample. Rather, models are constructed to identify non-uniform distribution (groupings) of samples (i.e. cluster analysis) or to explain the variance in the samples in terms of the observed features (e.g. principal component analysis and factor analysis). More recently, semi-supervised approaches have been developed to perform classification where only some of the samples have known class labels.

Supervised machine learning is generally carried out in two stages: training and testing. In the training stage, a model is constructed based on distinct patterns in training data associated with different classes. In the testing stage, the identified patterns are exploited to make predictions about future data.

The past several decades have seen significant advances in data availability, computing power, affordability of storage, and ease of data sharing. These changes, along with improvements in machine learning algorithms, have led to successful application of supervised machine learning across a wide variety of domains. In many cases, machine learning methods have performed tasks previously thought to be exceedingly difficult or impossible to automate (González-Reymúndez et al., 2017; Jagga & Gupta, 2014; Sun et al., 2008). Successful construction of machine learning-based models, however, generally requires an abundance of high-quality data. As the use of supervised machine learning becomes increasingly common, researchers have tried to identify mislabeled instances and correct their labels before building the model. It is important to note that these studies focus on detecting class mislabeling instead of individual mislabeling. Class mislabeling refers to the instance in which a sample's class (e.g. tumor vs healthy tissue) is labeled incorrectly, while individual mislabeling refers to a sample that is labeled as belonging to the wrong individual or source. Section 2.1 describes current work in identifying class mislabeling and Section 2.2 discusses research into identifying individual mislabeling.

2.1 Detecting Class Mislabeling

The problem of mislabeling is particularly concerning in supervised machine learning applications as labeled samples - which often come from human tissue - are sparse, and mislabeled instances constitute noise in model building. This would decrease the accuracy and reliability of the model. Several approaches have been explored for detecting class mislabeling with the aim of increasing data quality. The methods used in these studies can be categorized into two different types: classification and statistical approaches.

2.1.1 Classification Approaches

As mislabeled instances constitute noise in model building, their presence is reflected in the decrease of a model's classification performance. Different studies have utilized different metrics that define the classification performance of a model. Muhlenbach et al. (2004) proposed an algorithm to identify training noise that influences class separation with the aim of minimizing error rate of the classifier model. The authors projected each sample into a graph and determined edges that connect samples: two samples are connected if there are no other samples between them. The connections helped to determine if a dataset has good class separability, having a lower number of edges than a random graph that needed to be cut in order to obtain well-defined clusters (sub-graphs connected only by samples of the same class). Then, the sample's neighbors are examined. An instance is considered mislabeled if the majority of its neighbors are of different classes. The experiments were performed on a collection of ten domains from the UCI Repository of Machine Learning Databases[1]. The authors also investigated the optimum handling method for the identified suspected samples. They found that handling suspected samples via the schema of "relabelling or else removal" (relabel if a suspected sample's neighbors are of the same class, otherwise remove) yields a lower error rate in all datasets except Breast Cancer Dataset, where removing all suspected samples consistently yields the lowest error rate.

[1] https://archive.ics.uci.edu/ml/index.php

Sánchez et al. (2003) inspected different approaches that enhance the classification accuracy of Nearest Neighbor (NN) classifiers. The experiment was applied on five datasets from the UCI Repository of Machine Learning Databases[1]. Different approaches were performed on the dataset to filter out *bad* samples, samples that are mislabeled or outlier. The authors found that the depuration method (use leave-one-out method to predict a sample's class using k-NN classifier. The sample is relabeled if it has *k'* representatives among *k* neighbours, or removed otherwise) yields the best accuracy in 4 datasets: Liver, Pima, Cancer and Heart Datasets. This study is similar to the previous study (Muhlenbach et al., 2004), as both studies aim to reduce the noise of training data and improve model classification.

Venkataraman et al. (2004) developed a method for distinguishing between correctly labeled and mislabeled data sampled from video sequences. Instead of training several classifiers as in an ensemble-based method, one single classifier (SVM with a linear kernel) is trained on multiple representations of the data where each representation is built by different "discriminating" subspaces that are significant in class separation. Then leave-one-out (LOO) cross-validation is used to identify mislabeled data. Mislabeled data are those data which the annotated label is inconsistent with the predicted label. The mislabeled data were removed and the authors showed that removing the mislabeled data increased the LOO cross-validation accuracy overall.

The above studies focus on filtering noisy and atypical training samples to improve the quality of training data. The accuracy of relabeling is not evaluated independently.

Rather, overall classifier performance is the optimization objective. Furthermore, all these studies use datasets where the number of samples (*n*) is greater than the number of variables or features (*d*), a case that is not commonly seen in omics datasets (*d* is typically much greater than *n*). There are a few studies in detecting class mislabeling in bioinformatics data.

Malossini et al. (2006) proposed an algorithm that identified mislabeled samples by label perturbation and data misclassification. The algorithm iteratively perturbed the label of one instance within cancer microarray datasets and performed *n* iterations of LOO classification using an SVM with a linear kernel. An instance was identified to be mislabeled under two conditions, either: 1) the instance was consistently misclassified after the perturbation of other instances or, 2) perturbation of that instance resulting in improved prediction power of the resulting classifier. The first condition is reported to be a better strategy in identifying mislabeled samples and achieved an average precision of 0.67 and average recall of 0.92 on three real microarray datasets. One major drawback of this method is the long execution time as the method requires training of n^2 classifiers: *n* iterations of LOO classification for *n* iterations of perturbing the label of instances.

Knights et al. (2011) explored the identification of mislabeled samples solely based on classifier error rate. They trained prediction models (random forest and nearest shrunken centroid) and performed prediction on 16S rRNA microbiota data in two classification tasks (classifying general body habitats like skin vs gut, and classifying hand/keyboard samples by individual). False positives and false negatives were treated as mislabeled instances. They demonstrated that their algorithm was robust to noise and still able to predict correct labels, but only when the noise level is < 40% and the data exhibited clear separation between classes. The authors recognized that this approach will not be useful in a harder classification task where the data separation between classes is very subtle.

Martín-Merino (2013) proposed a similar algorithm to that of the previous study with two differences: the classifier model is built using SVM with a dissimilarity kernel and the datasets used are cancer microarray datasets. The sample and labels were mapped into feature spaces using the dissimilarity kernel and outliers were detected using one-class classification. The authors performed the algorithm on cancer microarray datasets and reported that the algorithm is more effective than a traditional SVM with a linear kernel. All of these studies can be characterized as outlier detection methods, in that they aim at removing noise to have a better separability between classes and achieve a better classification performance of models.

2.1.2 Statistical Approaches

In contrast with the studies in the previous subsection, some studies identify mislabeled instances by observing the statistical distribution of data. Westra et al. (2011) identified sample mislabeling by observing the deviation of gene expression z-scores in gene expression Quantitative Trait Loci (eQTL) datasets. The mean gene expression z-scores of different genotypes were computed. Significant cis-eQTLs, i.e., trait loci that have significant influence on expression of some specific genes, were first identified. Loci whose expression z-scores of those genes were highly distant from their genotype group mean were identified as mislabeled. The authors performed sample mislabeling checks on published datasets and found that four out of five datasets contain sample mislabeling. Overall, 3% of all samples were mislabeled and 15% more significant cis-eQTLs were identified after correction.

Lynch et al. (2012) took a similar approach, but they observed the misclassification of genotypes instead of the deviation of gene expression z-score. Similar to the previous study, significant cis-eQTLs were first identified. Then, the expression values of those significant genes were used to predict the genotype of samples. Any instances with inconsistent genotypes (predicted versus annotated) were identified as mislabeled.

Zych et al. (2017) utilized genotype perturbation and identified mislabeled samples by observing changes in the t-statistic value. The rationale behind this approach is that if a mislabeled genotype is perturbed to its true label, the overall t-statistic value between different genotypes would increase and vice-versa. The algorithm achieved an area under the curve (AUC) of 0.8 to 1.0, depending on genetic similarity of datasets (the more dissimilar the dataset, the higher the AUC achieved). The authors performed the algorithm on public worm gene expression datasets (Snoek et al., 2012; van der Velde et al., 2013) and 1.9% (4 / 208) of *C. elegans* recombinant cell lines are found to be mislabeled. One drawback of the algorithm is a high execution time exacerbated by the perturbation of genotype. Each perturbation requires a new calculation of t-statistics for every genotypes-gene pair. The method requires high performance computing when a dataset has a fairly large set of samples. The worm datasets used in the studies have ~120 single nucleotide polymorphisms (SNP) but humans have ten of millions. This limits the scalability of the algorithm.

As with the previously-described classification-based approaches, all of the statistical approaches mentioned above identified mislabeled samples with the aim of assuring the correctness of class labels. Further investigation into the source of error is usually not performed. However, detecting class mislabeling is not sufficient in the field of precision medicine, where a data instance should not only be correctly assigned to its class but also attributed to the correct patient.

2.2 Detecting Individual Mislabeling

Very few studies focus on detecting individual mislabeling. Broman et al. (2015) identified mislabeled samples by inspecting the concordance of gene expression data across different tissue types. The datasets are generated from six tissue types from the same population of mice. Every mouse subject had six tissues (adipose, gastrocnemius muscle, hypothalamus, pancreatic islets, kidney, and liver) extracted and sequenced using an Affymetrix microarray platform. The rationale of the approach is that the concordance of gene expression between two tissues from the same mouse should be high. While the authors were able to identify and correct mislabeled samples to the individual level, the study focuses on transcriptomics data from several tissues. The method is not directly applicable to study designs where only one type of tissue is sampled for different omics. In addition, the approach does not provide a mechanism for automated correction. Rather, it relies upon manual intervention to identify mislabeled samples.

3. MATERIALS AND METHODS

3.1 Data Collection

Three multi omics datasets were collected from National Institute of Health’s Clinical Proteomic Tumor Analysis Consortium (CPTAC)[2]. These datasets were generated for three different cancers: colorectal, kidney and lung. Intensive manual inspection has been performed and there was no observed data mislabeling upon publication.

3.1.1 Colorectal Cancer Dataset (COAD)

The colorectal cancer dataset was merged from two colon rectal cancer cohorts, 85 from Zhang et al. (2014) and 96 from Vasaikar et al. (2019). Two types of omics data were collected: transcriptomics and proteomics. Expression level of mRNA was quantified based on Fragments Per Kilobase of transcript per Million mapped reads (FPKM). Protein fragmentation and sequencing were performed through Liquid Chromatography with Mass Spectrometry (LC-MS/MS) and protein abundance was measured based on spectral counting (the total number of MS/MS spectra acquired for peptides from a given protein). For both proteomics and RNA-seq data, genes with more than 50% missing values were removed, except for genes located in X or Y chromosomes. The missing values were imputed using Random-Forest based imputation (Stekhoven & Bühlmann, 2011) except

[2] https://proteomics.cancer.gov/programs/cptac

for sex chromosome genes where the missing values were replaced by zero. This resulted in a total of 17220 gene and 4105 protein features. The proteomics data were then normalized using quantile normalization whereas the RNA-seq data was normalized using the trimmed mean of M-values normalization method (TMM) (Robinson & Oshlack, 2010). Since the dataset was integrated from two cohorts, batch correction was performed on both proteomics and RNA-seq data using Combat (Johnson et al., 2007) after data normalization. Quality control analysis was performed using metaX (Wen et al., 2017) before and after batch correction.

3.1.2 Kidney Cancer Dataset (CCRCC)

The kidney cancer dataset (Clear Cell Renal Cell Carcinoma) was collected from Clark et al. (2019). Two types of omics data were collected: transcriptomics and proteomics. The expression level of mRNA was quantified based on FPKM while the expression level of protein was measured based on spectral counting after performing the MS/MS pipeline. The samples were manually inspected for any sampling error and one sample was removed due to low self correlation between RNA and Protein profiles (Clark et al., 2019). The features with missing rate > 50% were filtered and proteomic missing values were imputed using DreamAI[3], an ensemble algorithm developed during the National Cancer Institute-Clinical Proteomic Tumor Analysis Consortium (NCI-CPTAC) Dream Proteomics

[3] https://github.com/WangLab-MSSM/DreamAI

Imputation Challenge[4]. There were a total of 19275 RNA features and 10127 protein features. Lastly, the mRNA expression levels and global protein abundances were normalized to a standard normal distribution.

3.1.3 Lung Cancer Dataset (LUAD)

The lung adenocarcinoma dataset was collected from Gillette et al. (2020). Four types of omics data were collected: transcriptomics, proteomics, copy number variation (CNV) and microRNA (miRNA). The RNA transcript read counts were upper-quartile normalized and transformed into Reads Per Kilobase of transcript, per Million mapped reads (RPKMs). Protein abundance was quantified based on spectral counting and normalized TMT ratios. The features with missing rate > 50% were filtered and proteomic missing values were imputed using the DreamAI tool[5]. CNV analysis was performed using CNVEX[6], an algorithm which uses several probabilistic and optimization algorithms to estimate the copy number from whole genome sequencing (WGS) and whole exome sequencing (WES) data. Expression of miRNA was quantified using a variant of the small RNA quantification pipeline developed for TCGA (Chu et al., 2016). The number of features in each omics dataset are: 19275 (RNAseq), 7556 (proteomics), 19817 (CNV), and 1881 (miRNA).

[4] https://www.synapse.org/#!Synapse:syn8228304/wiki/413428

[5] https://github.com/WangLab-MSSM/DreamAI

[6] https://github.com/mctp/cnvex

3.2 Mislabeling Simulation

Three mislabeling error patterns were observed in various TCGA or CPTAC datasets: swapping, duplication and shifting (Clark et al., 2019). These similar error patterns were introduced into the datasets during simulation and the simulation mechanisms were described as below.

3.2.1 Swapping

Swapping errors occur when the patient labels of two samples from different subjects are swapped. Swapping errors can occur in any type of omics data and were simulated by swapping the data of two samples.

3.2.2 Duplication

Duplication error is the replication of data from one patient. This may be an electronic duplication of the data or, more frequently, when a tissue sample is divided and unintentionally assayed multiple times. The resulting duplicate data replaces the data for another sample. This is often a sample associated with a different subject. To simulate a duplicate data that accurately reflects real duplicates, actual proteomics replicates were referred to. There are two additional actual proteomics replicates in the COAD dataset and the replicates were found to have Pearson correlation coefficient > 0.9 with their original counterparts. To simulate a duplicate data, the original data was added with a perturbation equal to the standard deviation of each gene *i* as in $Sample(i)_{duplicate} = Sample(i) \pm \sigma/\alpha$, where σ is a standard deviation of the gene *i* and α is a scale factor for the σ. For each gene, the perturbation is either added to or removed from (randomly selected) the original value.

The scale factor was empirically optimized to yield correlation coefficients between simulated duplicates similar to that observed in actual proteomics duplicates. The changes of score difference with respect to the changes of scale factors are visualized in Figure 3.1. Score difference is the difference of sample correlation with itself and the average of sample correlation with others. The original data have a score difference of 0.495. A low α increases the perturbation variance and decreases the score difference, indicating that the duplicated data is more similar with random samples (no differences between self and other sample), whereas a high α increases the score difference until it reaches the level of original data (no differences with original sample). It was found that a scale factor of α = 1.0 (pivot of the elbow line in Figure 3.1 middle) resulted in a correlation coefficient > 0.9 between simulated RNAseq replicates and the original samples. The original data of another randomly selected patient was discarded and displaced by the simulated data.

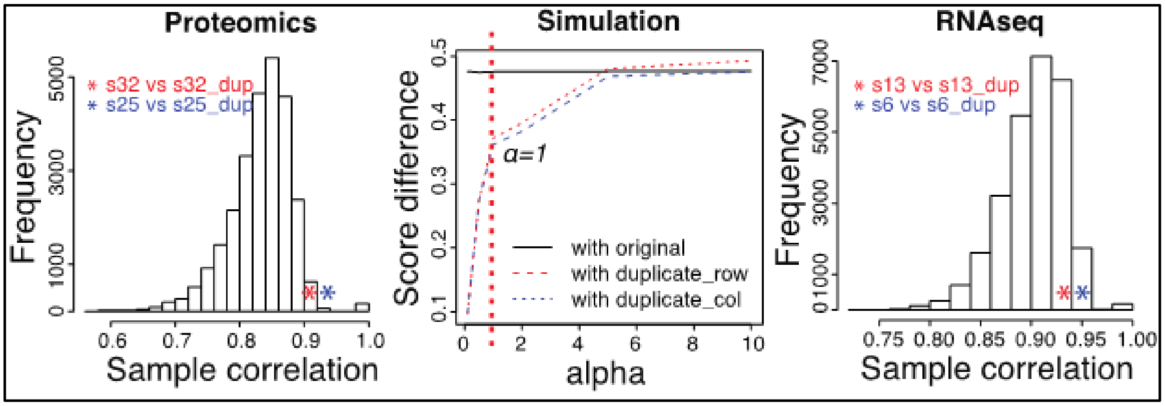


Figure 3.1: Simulation of duplicated samples. (left) The actual proteomic replicates have high sample correlation and this experimental data is used as a reference for simulation. (middle) Simulation of the scale factor α to control sample similarity score between RNA-seq and proteomics. Two dashed colored lines mark the sample simulation for RNA-seq and proteomics respectively. (right) The scale factor α = 1 is chosen to simulate RNA-seq duplicated samples which have similar sample correlation with the reference in proteomics dataset (Figure adapted from Yoo et al., 2021).

3.2.3 Shifting

Shifting errors indicate the displacement of several samples to another sample in a sequential manner (A to B, B to C, and C to D). Shifting errors can occur in any type of omics data and one shifting event always involves several samples, typically ranging from 3 to 6.

3.3 Pearson Correlation Coefficient

Pearson Correlation Coefficient measures the degree of relatedness between two sets of data. It measures the linear relationship between them and is the ratio between the covariance of two variables and the product of their standard deviations (Equation 1 and 2).

$$\rho X, Y = \frac{\sigma XY}{\sigma X \sigma Y} \qquad (1)$$

$$\sigma XY = \sum (x_i - \overline{x})(y_i - \overline{y}) \qquad (2)$$

Where ρX,Y = correlation of X and Y, σXY = covariance of X and Y, σX = standard deviation of X, σY = standard deviation of Y.

When computing gene correlation across two omics data, X represents the expression values of a specific gene in one omics data while Y represents the expression values of a specific gene in another omics data with the same sample order as X; whereas when computing sample correlation across two omics data, X represents the expression values of genes of a sample in one omics data while Y represents the expression values of the genes of a sample in another omics data with the same gene order as X.

3.4 Stable Matching Algorithm

Given a sample correlation heatmap which indicates the correlation of two samples from two different omics data, the challenge is to find a set of matchings that each matching pairs two samples with the highest correlation as possible. As such, we employed the Gale-Shapley algorithm for finding a solution to this stable matching problem (Gale & Shapley, 1962). The input of the algorithm is lists of preferential rank, one for each sample, ranking from the highest Pearson correlation to the lowest. The preferential ranks were generated from the sample correlation heatmap, where the ranking is done in each row for the respective samples from omics x and in each column for the respective samples from omics y. The output is a set of matchings, pairing two samples from different omics data.

Below is the pseudocode of the stable matching algorithm used in this thesis:

```
Input: Preferential ranks for every samples
Output: A set of matchings pairing two samples, one from omics x and one from omics y

Initialize m ∈ samples from omics x and w ∈ samples from omics y to be unpaired
While ∃ m which is not paired with some w:
    w := first sample of omics y on m's list to which m has not yet tried to pair
    if ∃ some pair (m', w) then
        if w has upper rank with m to m' then
            m' becomes unpaired
            (m, w) become paired
        end if
    else
            (m, w) become paired
    end if
repeat
```

3.5 Evaluation

To evaluate the performance of the proposed approach in detecting and correcting mislabeling, several simulation experiments were performed. In every simulation experiment, the true datasets were used to simulate artificial datasets with mislabeling errors as described in the previous subsection. The true individual labels of the simulated datasets remained hidden and were used for evaluation later. The expected mislabeling rate in real life datasets is low, ranging from 0-20%. Thus, we evaluated the proposed approach using F1 scores (Equation 3, 4, and 5), the harmonic mean of precision and recall. The proposed approach should be able to detect as many mislabeled samples as possible without overcorrecting those correctly-labeled samples.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \qquad (3)$$

$$precision = \frac{TP}{TP + FP} \qquad (4)$$

$$recall = \frac{TP}{TP + FN} \qquad (5)$$

True positive (TP) is the number of positives correctly identified, true negative (TN) is the number of negatives correctly identified, false positive (FP) is the number of negatives incorrectly identified as positive, and false negative (FN) is the number of positives incorrectly identified as negatives. Three levels of F1 scores were used in evaluation: sample level, data level, and correction level. Each level has different criteria to consider an instance as a true positive, with the subsequent level having stricter criterias than the previous.

Sample level F1 score evaluates the performance of detecting samples with mislabeled data. A positive instance is the sample with mislabeled data regardless of omics type. If the corrected labels of any omics data do not match the original sample label, it is considered a mislabel identification at the sample level. Data level F1 score evaluates the performance of identifying correctly the types of mislabeled data. A true positive instance is an instance in which all the corrected labels of omics data match the original sample label except the mislabeled one. Correction level F1 score evaluates the performance of correcting the individual label. The corrected labels of all omics data should match exactly the true labels to be considered as a true positive.

4 MISLABELING DETECTION

This approach is based on the rationale that multiple types of omics data characterized from the same patient have intrinsic relationship between each other and could be utilized to extract a signal for aligning omics data. The signal should possess two characteristics: 1) it should be general such that it could be extracted from every sample, but at the same time, 2) it should be highly specific in every individual such that every sample has a strong signal to itself but not to other individuals.

The aim of this chapter is to investigate if such a signal could be extracted from different omics data and how accurate the signal would be in detecting mislabeled samples.

4.1 Pairwise Alignment

4.1.1 Correlation Signal Extraction

Central Dogma of molecular biology describes the flow of genetic information from DNA to RNA and from RNA to protein via the processes of transcription and translation. The copy number of a gene, expression level of a gene transcript and expression level of a gene product are correlated to some extent. The correlation between these three omics data could be exploited to determine whether two data collections from different types of omics assays belong to the same patient. We employ the following procedure to extract the correlation signal that could accurately align two different omics data from the same patient.

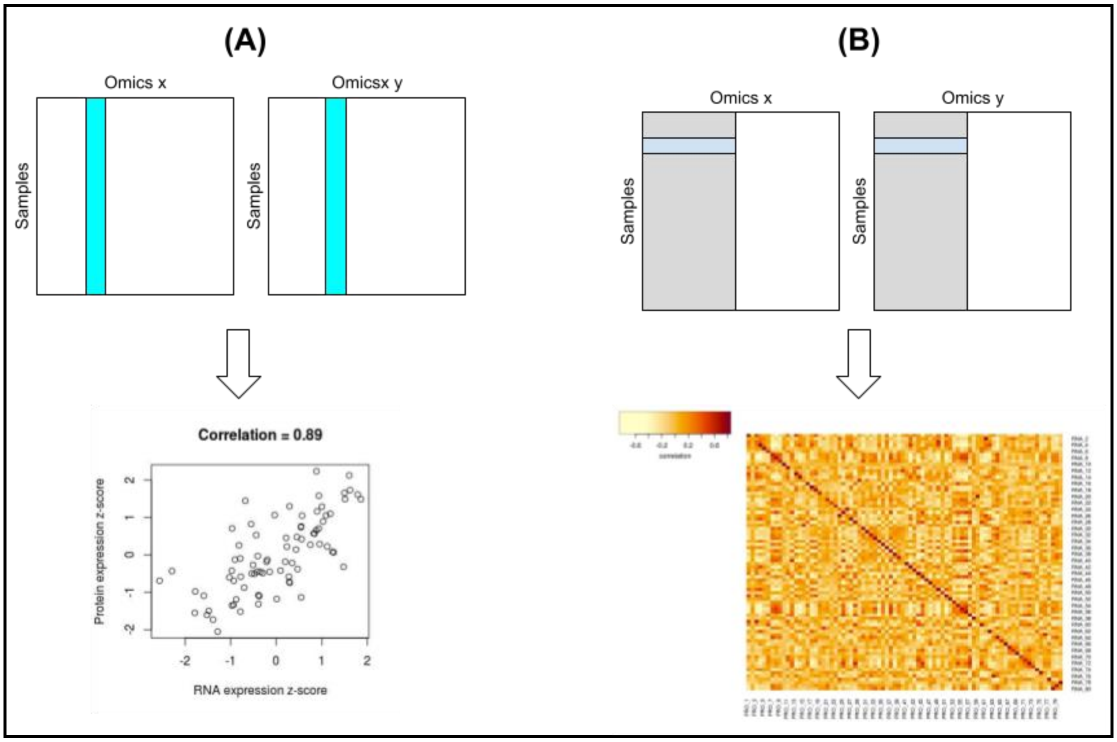


Figure 4.1: Imputation of sample correlation for performing pairwise alignment.

First, gene correlation between two omics data is computed and genes with high correlation (cor > 0.5) are extracted. The expression value of these highly correlated genes are then used to compute sample correlation between two omics data, generating a sample correlation matrix, C with a dimension of $N \times N$.

4.1.2 Coexpression Signal Extraction for miRNA Data

It is possible to extract correlation signals between RNA-seq, proteomics and CNV data as these omics have the same gene features. However, this is not the case for microRNA (miRNA). MicroRNA refers to a short single-stranded RNA (~22 nt) molecule and mediates RNA silencing through base pair pairing. MicroRNA is not protein coding and thus, does not have the same gene features with other omics data.

Instead of correlation signal, the coexpression signal is utilized to align miRNA samples. Around 70% of mammalian miRNA are embedded within a host gene and these miRNA are known as intragenic miRNA (Rodriguez et al., 2004). These miRNA were found to share a common transcription unit with their host genes and always co-transcribe together (Baskerville & Bartel, 2005; Ramalingam et al., 2014). The coexpression patterns were investigated in this chapter to determine their utility in miRNA sample alignment.

To pair each miRNA with its host gene, the annotated human genome build dataset (GRCh38.p13) was downloaded from the NCBI Ensembl[7] website and the annotated miRNA dataset (release v22) was downloaded from mirBase[8]. The annotated miRNA dataset aligns to the human genome of same build GRCh38 and contains a total of 1919 unique miRNA entries. If the genomic position of a miRNA is within the genomic range of a gene, then they are considered as a miRNA-gene pair. A total of 1647 miRNA-gene pairs were extracted, comprising 1419 unique miRNAs and 1173 unique genes.

For every pair of miRNA and its host gene, the feature correlation was imputed between miRNA with three other omics data (RNAseq, proteomics and CNV). Those miRNA-gene pairs that are highly correlated were extracted. The expression value of these

[7] http://useast.ensembl.org/index.html

[8] http://www.mirbase.org/

highly correlated genes were then used to compute sample correlation, generating a sample correlation matrix, C with a dimension of $N \times N$.

4.1.3 Stable Matching for Detection

The correlation matrix C, was used as the preferential ranking for a stable matching algorithm, with the ranking being ordered by the correlation descendingly. The stable matching algorithm outputs N pairs of matching sample pairs with matching scores, the sum of preferential ranks of both omics data towards each other. Ideally, omics data from the same patient should have the top rank with each other, contributing to a matching score of 2.

If a sample pair consists of omics data of different patients, those are considered as mislabeled samples. The stable matching algorithm pairs exactly one-to-one omics data thus data that are left out (due to duplication) will be paired despite having very low correlation signals with each other. Thus, a sample pair with a matching score > N/10 were also considered mislabeled.

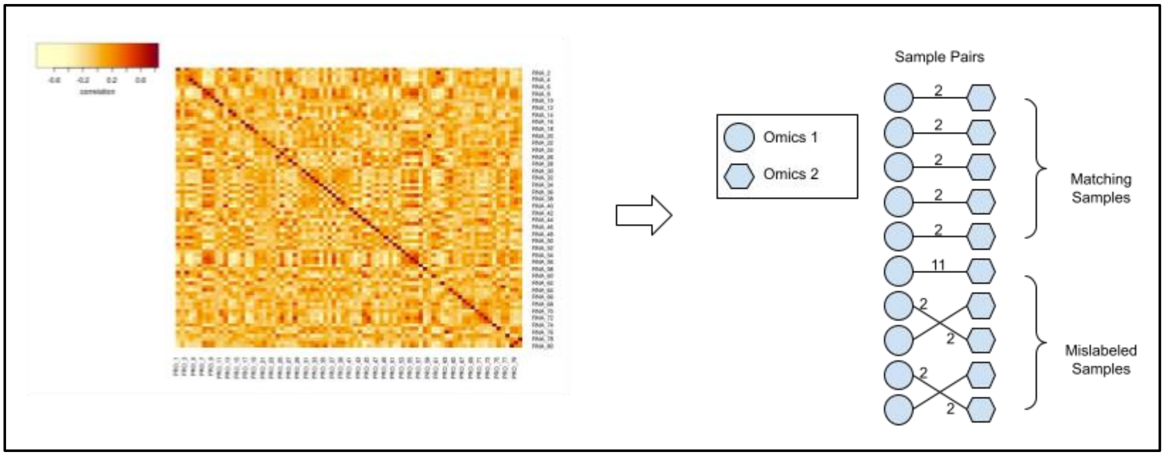


Figure 4.2: Stable Matching algorithm pairs every instance from two different omics data.

4.2 Evaluation

4.2.1 Gene Correlation Inspection

The distribution of genes Pearson Correlation between different omics was inspected. In Figure 4.3, the gene correlations follow a normal distribution and have a mean ranging from 0.24 to 0.53. This implies that there is a reasonable correlation between these omics data to be used for pairwise alignments.

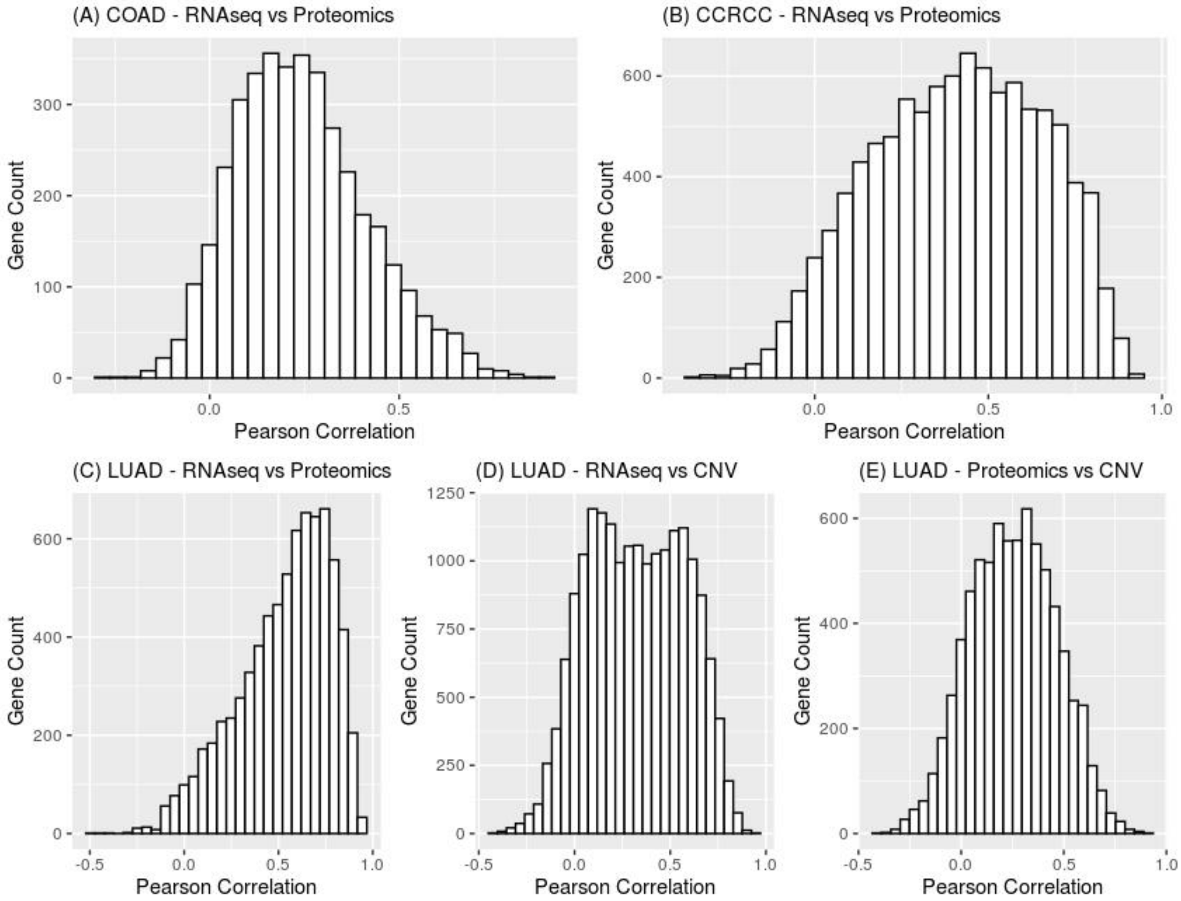


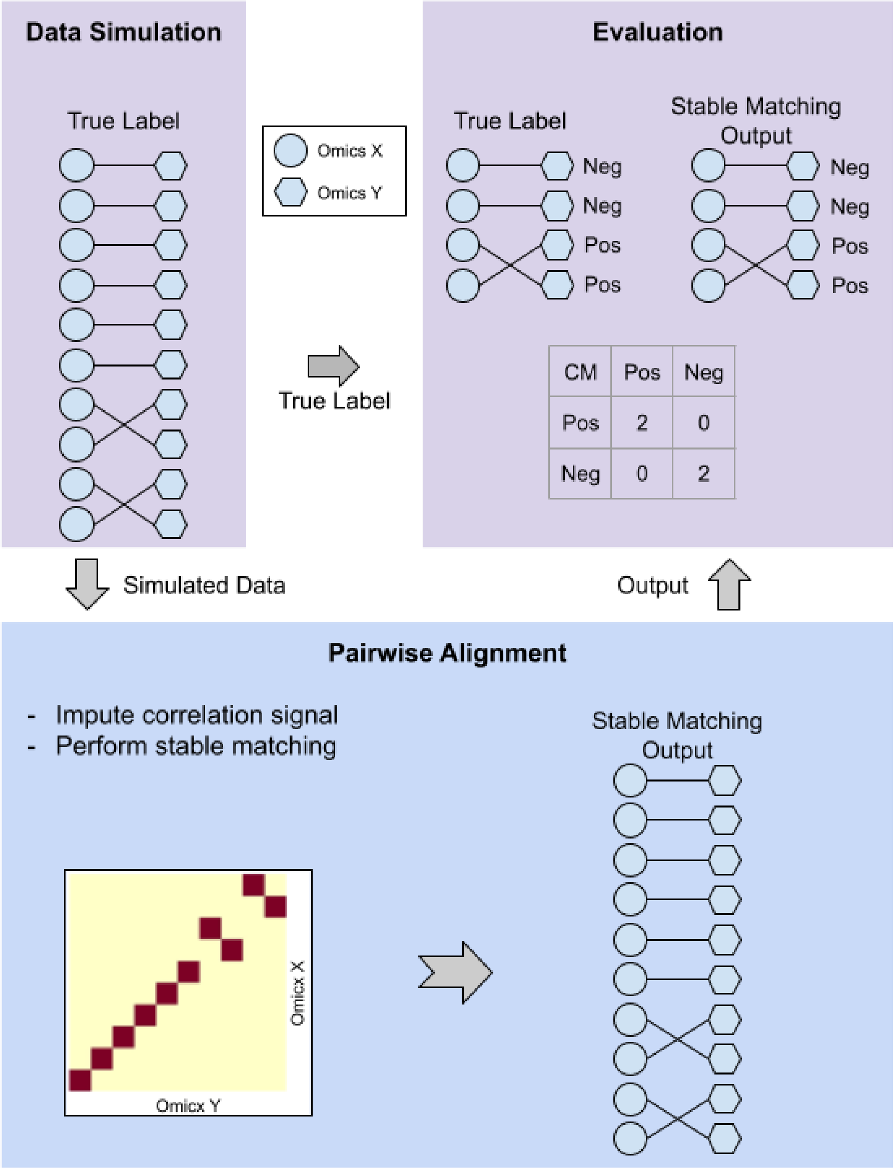
Figure 4.3: Distribution of Gene Pearson Correlation between different pairs of omics datasets: (A) Colon adenocarcinoma RNAseq and Proteomics data, (B) Clear Cell Renal Cell Carcinoma RNAseq and Proteomics data, (C, D, E) Lung Adenocarcinoma RNAseq, Proteomics and CNV data

Dataset	**CPTAC COAD**	**CPTAC CCRCC**	**CPTAC LUAD**		
Cancer	Colon Adenocarcinoma	Clear Cell Renal Cell Carcinoma	Lung Adenocarcinoma		
Omics Data (Number of Features)	RNAseq (13172)	RNAseq (19275)	RNAseq (19275)	RNAseq (19275)	Proteomics (7556)
Omics Data (Number of Features)	Proteomics (4105)	Proteomics (10127)	Proteomics (7556)	CNV (19817)	CNV (19817)
Number of Overlapped Features	3866	9946	7416	18707	7510
Mean of Pearson Correlation	0.2422	0.4099	0.5315	0.3233	0.2538

Table 4.1: Mean of gene correlation between two omics dataset

4.2.2 Simulation Experiments

To investigate if the correlation signal extracted could be a useful indicator to inform mislabeling, the original datasets were used to simulate mislabeled datasets by the process of bootstrapping and artificial mislabeling. In each simulation, a fixed number of omics instances (100 in COAD; 80 in CCRCC and 80 in LUAD) were randomly selected and mislabeling errors with an error rate of *e* were introduced as described in Section 3.2. The process was repeated 10 times for each error rate, *e* = [0.1, 1.0].



CM	Pos	Neg
Pos	2	0
Neg	0	2

Figure 4.4: Simulation experiment for evaluating the performance of correlation signal in detecting mislabeled samples. The process was repeated 10 times in each pair of omics data for each error rate.

The pairwise alignment algorithm was performed on each simulated dataset. Every patient whose omics data is mislabeled, is a positive instance and only a successful detection of the patient is treated as a true positive. The complementary principles applied to a true negative instance. The sample level F1 score is obtained in each simulation and shown in Figure 4.5 (A).

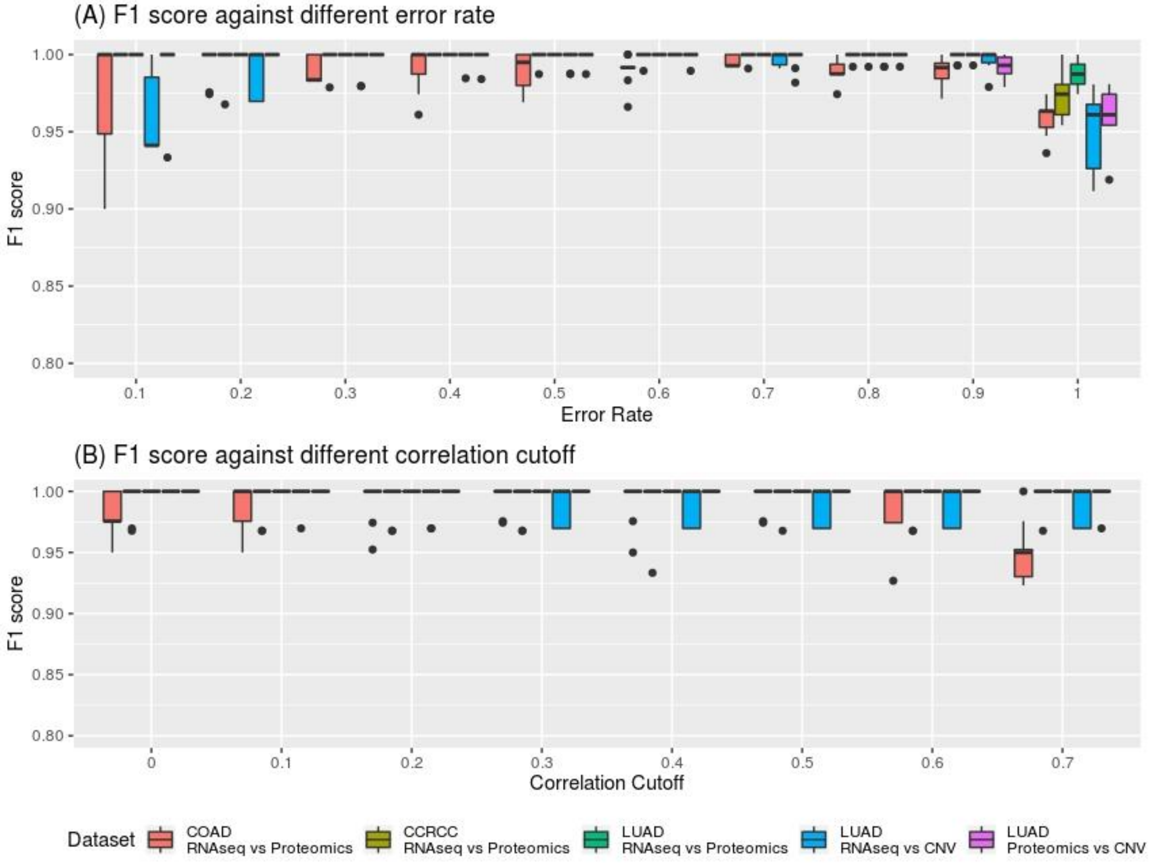


Figure 4.5: Sample level F1 scores of mislabeling detection between different pairs of omics datasets: (A) across different error rates using default correlation cutoff of 0.5, (B) across different correlation cutoff with fixed error rate of 0.2.

Figure 4.5 (A) shows that the correlation signal could accurately align samples across different datasets and different pairs of omics data, the average F1 score across all the datasets is 0.99. The algorithm was shown to be robust against error rate and achieved F1 score > 0.97 across different error rates. It is counterintuitive that the algorithm could achieve high F1 scores when the error rate = 1.0. This is because when the error rate = 1.0, the data do not have any useful correlation and the sample is paired randomly. Thus, all the samples were detected as mislabeled because of random pairing that does not yield any matching pairs. The simulation experiments were also repeated with fixed error rate, e = 0.2 but with different correlation cutoffs. This is to determine the optimum correlation cutoff to extract gene features. Figure 4.5 (B) shows that the algorithm is robust against different cutoffs and the F1 scores were ~0.94 across different cutoffs.

4.2.3 miRNA Coexpression Inspection

Several publications showed that miRNA always coexpress with its host genes (Baskerville & Bartel, 2005; Ramalingam et al., 2014). The miRNA data of the lung adenocarcinoma dataset (N = 107) was inspected and the distribution of miRNA-gene pairs Pearson correlation was plotted (Figure 4.6). The histograms show that miRNA has reasonable coexpression pattern to all the omics data, with the strongest correlation to RNAseq data with a mean of 0.231 while the lowest correlation to CNV data with a mean of 0.106.

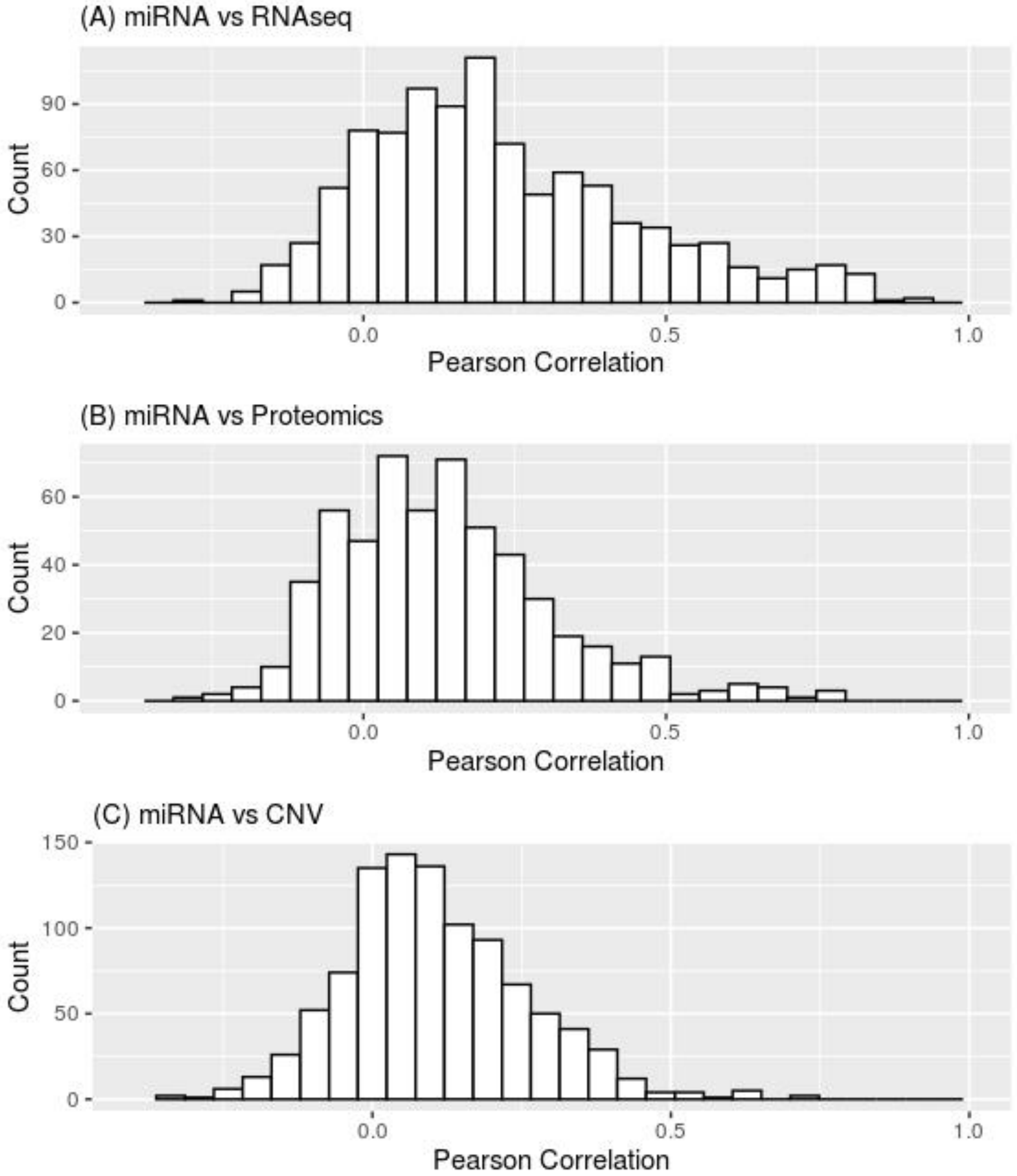


Figure 4.6: Distribution of Pearson correlation of miRNA-gene pairs between (A) miRNA with RNAseq, (B) miRNA with proteomics, and (C) miRNA with CNV data of the Lung Adenocarcinoma (LUAD) dataset.

Alignments of omics	Number of pairs of miRNA and host gene	Mean of Pearson Correlation
miRNA with RNAseq	1166	0.2305
miRNA with proteomics	646	0.1331
miRNA with CNV	1182	0.1055

Table 4.2: Number of miRNA-gene pairs and the mean of Pearson correlation.

4.2.4 Simulation Experiments for miRNA Data

Simulation experiments were performed with three different pairs of omics data with miRNA: miRNA to RNA, miRNA to Proteomics, and miRNA to CNV. The original datasets were used to simulate mislabeled datasets (*N* = 100) with an error rate of 0.1. The process was repeated 50 times and the pairwise alignment was performed on simulated datasets iteratively using different correlation cutoffs. The performance of the pairwise alignment algorithm in aligning miRNA data was inspected.

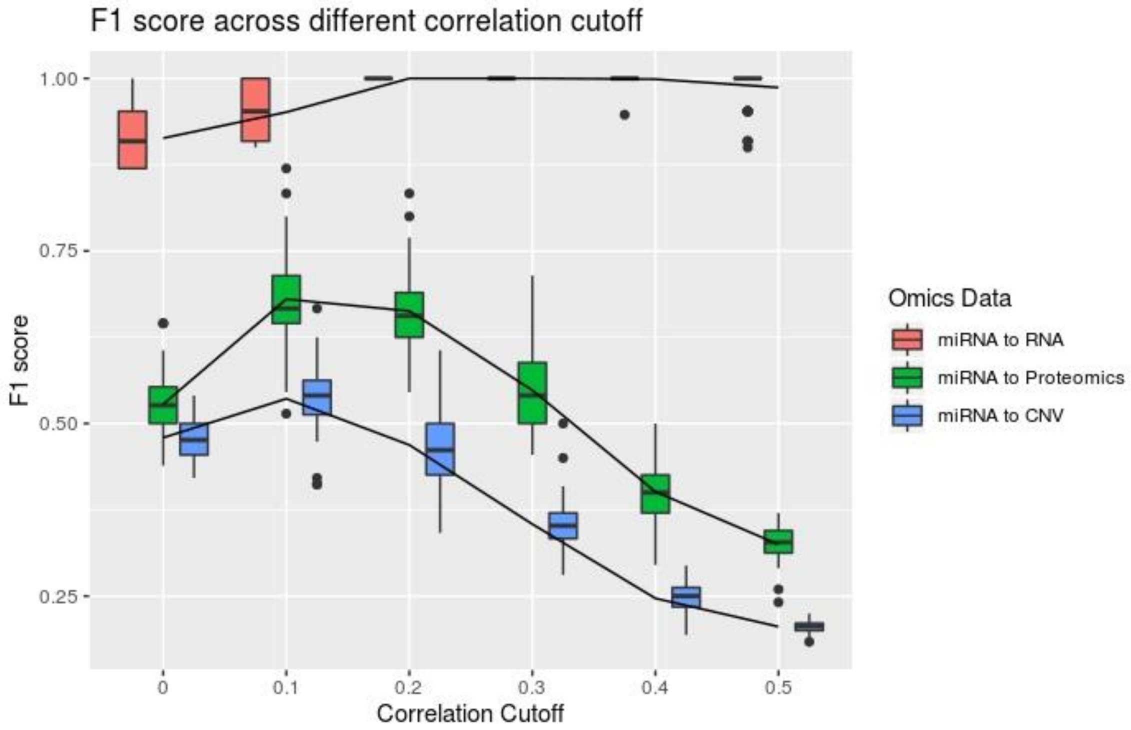


Figure 4.7: Sample level F1 scores of detecting mislabeled samples in pairwise alignment of miRNA data with three other omics data.

Figure 4.7 shows that the algorithm achieved an F1 score of 1 when aligning miRNA with RNAseq data, whereas aligning miRNA with Proteomics data achieved the highest F1 score of 0.68 and aligning miRNA with CNV data achieved the highest F1 score of 0.54. The coexpression signal is proved to be useful in aligning miRNA samples with RNA samples, but not with proteomics nor CNV data.

It is hypothesized the reason for the low F1 score in aligning miRNA to Proteomics or CNV samples is due to low number of correlated feature pairs. To validate the hypothesis, the simulation experiments were repeated but in each simulation, a different number of correlated feature pairs (corr > 0.1 or 0.2) were extracted randomly to compute sample correlation for pairwise alignment. Figure 4.8 shows that the higher the number of feature pairs, the higher the F1 score achieved. However, the F1 score was limited by the number of correlated feature pairs in Proteomics and CNV data. RNAseq data has the highest coexpression signal with miRNA data, and yet it requires at least 300 feature pairs (corr > 0.2) to achieve an F1 score of 1. Proteomics and CNV data do not have sufficient correlated feature pairs (corr > 0.2) to miRNA data, which normally capped at around 125 and 175 features pairs. Combined with the weaker coexpression of these datas to miRNA data, the F1 score achieved is not sufficient for accurate label prediction. Yet, the increasing trend of F1 score against the number of features, suggests that a higher F1 score, and thus a useful correlation signal could be achieved if there were more feature pairs available. One surprising finding is that the pipeline has high recall regardless of the number of feature pairs and the low F1 score is due to low precision. This indicates that all mislabeled samples are being identified and a matching sample can be treated as correctly labeled with confidence.

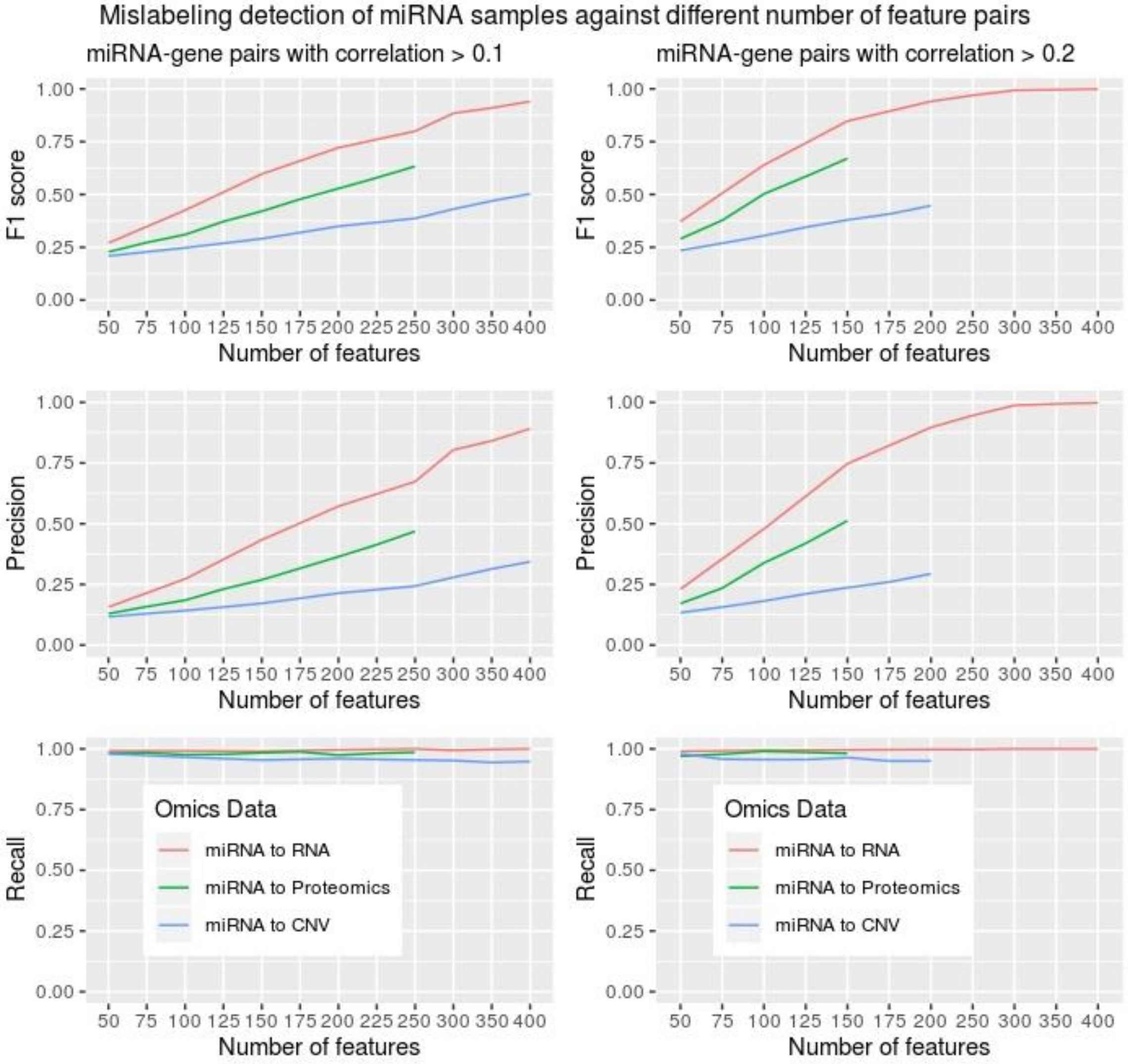


Figure 4.8: Sample level F1 score achieved against different number of feature pairs. The feature pairs were selected randomly from a set of features with correlation > 0.1 (left) or 0.2 (right). Higher number of correlated feature pairs achieved a higher F1 score, but it is limited in Proteomics and CNV data.

4.3 Discussion

In this chapter, an algorithm was proposed to extract correlation signals between two omics data and use the signal to perform pairwise alignments of samples. The results show that the correlation signal is reliable and achieved a high F1 score (> 0.95) in detecting mislabeled samples across RNAseq, Proteomics and CNV data from three cancer datasets. The detection algorithms are robust against different error rates and achieved average F1 scores > 0.95 even in a dataset with high error rate.

The coexpression signal extracted from miRNA with RNAseq data was also shown to be an accurate indicator to align miRNA and RNAseq samples, with an F1 score of 1.0. However, alignment of miRNA to another two omics samples did not achieve satisfactory F1 scores (0.68 to proteomics; 0.54 to CNV). This is due to the low number of correlated feature pairs between these omics data. Higher number of feature pairs are required to extract reliable coexpression signals that could accurately align proteomics and CNV samples. It is estimated that Proteomics data requires at least 450 pairs while CNV data requires at least 800 pairs.

5 MISLABELING CORRECTION IN TWO OMICS DATA

The preceding chapters show that pairwise alignment across omics data sources can accurately detect mislabeled samples in multi omics experiments. However, alignment alone is insufficient for correcting the identified mislabelings. To correct the label, it is important to investigate the source of error to determine which omics data get mislabeled. Here we propose an algorithm that utilizes the class label to determine the source of error. The class label could be any clinical attribute collected in the dataset. In this chapter, the aim is to investigate the accuracy obtained by using various clinical attributes as the class label in correcting mislabeled samples.

5.1 Attribute Prediction

Other than omics data characterization, clinical attributes data of the patients are often collected in a cohort study. Different studies collect different clinical attributes of the patients, but sex is one of the most common attributes across all the studies. Thus, sex phenotype and genotype was chosen to be utilized in this work for label correction.

Due to the dimensionality curse in bioinformatics data, two feature selection methods were used. The first is the selection of sex-linked genes in sex chromosomes. The missing values in sex-linked genes were replaced by 0 as they are assumed to be either absent (i.e., the absence of the Y chromosome in females) or repressed (i.e., X chromosome inactivation in females). The second method is elasticNet regularization during the model training. One classifier model was trained for each omics data and the model used in this work is a regularized weighted logistic regression model. Class weighting was used to compensate for any class imbalance issue and logistic regression (LR) model was chosen due to its simplicity and efficiency in training.

There is an inherent circular problem in training a model to predict gender from multi omics data from a particular study, given the possibility that sample mislabeling may have occurred in the training data. To resolve this issue, the model is trained in two steps. First, matched samples from the study data are used to train a model using k-fold cross validation (CV). Next, samples for which sex genotype is mispredicted during testing folds are identified as suspected cases of mislabeling, and are excluded from the model. Next, the model is retrained with only high-confidence non-mislabeled samples from the study. Finally, the re-trained model is used to predict sex genotype across all samples.

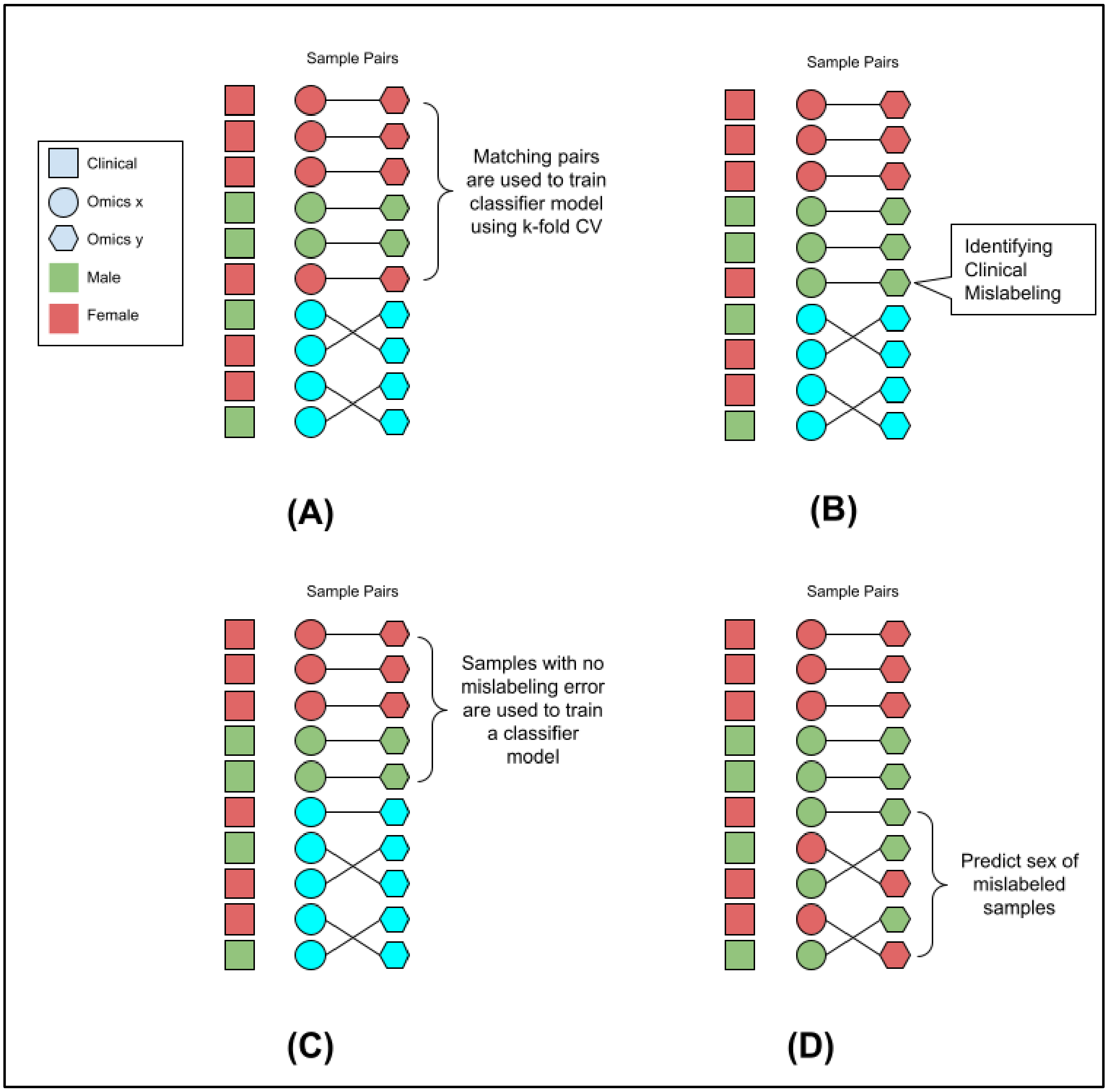

Clinical
Omics x
Omics y
Male
Female
Sample Pairs
Matching pairs are used to train classifier model using k-fold CV
(A)
Sample Pairs
Identifying Clinical Mislabeling
(B)
Sample Pairs
Samples with no mislabeling error are used to train a classifier model
(C)
Sample Pairs
Predict sex of mislabeled samples
(D)

Figure 5.1: The workflow of predicting sex label of omics data: (A) Matched samples from the study are used in the first round of model training using k-fold cross validation (B) Samples which sex genotype is mispredicted were identified (C) Suspected mislabeling case is excluded from second round of model training (D) Trained model is used to predict sex genotype across all samples.

5.2 Individual Label Correction

With the detected mislabeled samples and predicted sex label, an automated correction algorithm was proposed. The algorithm inspects pairwise alignment patterns along with the predicted sex label to correct the label. There are three types of mislabeling error and the algorithm corrects each type of error with different mechanisms.

Swapping errors are characterized by the cross alignments between two patients' samples. To determine which omics data get swapped, the predicted sex labels of omics data are checked against the annotated sex labels. Each checking generates an error rate, the difference between annotated label and predicted probability. The omics data with a higher error rate is determined to be mislabeled and the labels will be corrected.

The identification of duplication error is complicated, as the stable matching algorithm pairs the samples in a strict one-to-one manner; a duplicated sample will not pair with its matching sample. Hence, the identification of duplication cases relies on the matching score. Due to the displacement of duplicated data, there is another sample with no matching sample and will always spuriously paired with the duplicated data despite having low correlation. Due to the low correlation, this pairing will have a high matching score. The highest possible matching score is $2N$ and a threshold of 5% ($2N/20$) is used, assuming that the spurious pairing will only have a 5% random chance to have a low matching score. Thus, we used $N/10$ as the threshold. If a sample pair has a matching score higher than the threshold, it is suspected to be a duplication case. To determine which omics data get duplicated, the algorithm inspected the highest correlation each data has with the other samples. The duplicated data is supposed to have high correlation with its sample, and thus the one with the higher correlation has its label corrected.

Shifting cases always start with a duplication event. Before correcting the label, the algorithm first identifies the shifting chain. The shifting chain starts with a duplicated sample, which is identified as 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 matching score higher than the threshold $N/10$. After the shifting chain is identified, the next step is to determine which omics data get shifted. The predicted sex labels of omics data are checked against the annotated sex labels. This checking step is the same as checking swapping errors but with several samples in the shifting chain. Each checking generates an error rate, the difference between annotated label and predicted probability. The omics data with a higher error rate is determined to be mislabeled and the labels will be corrected.

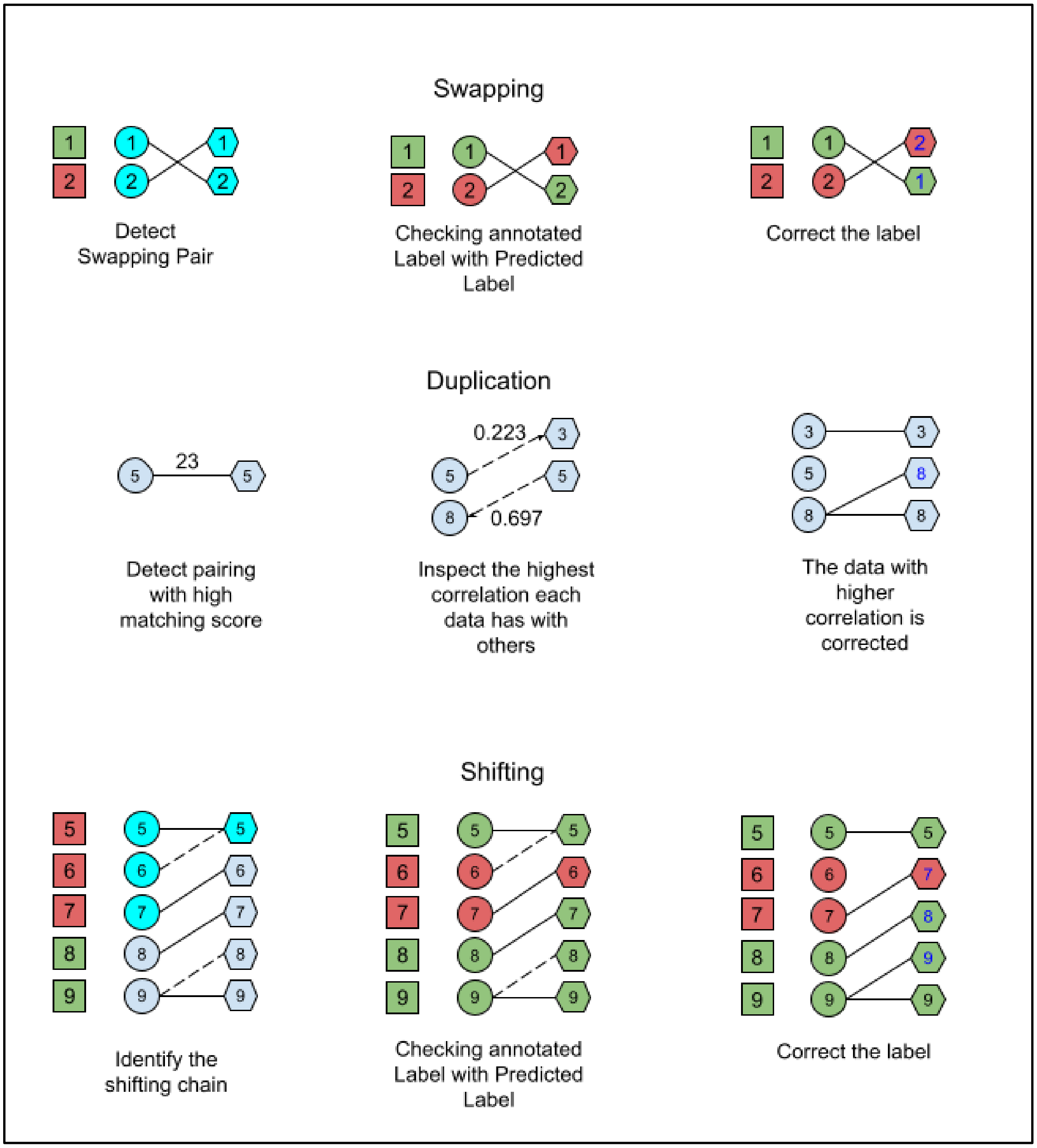
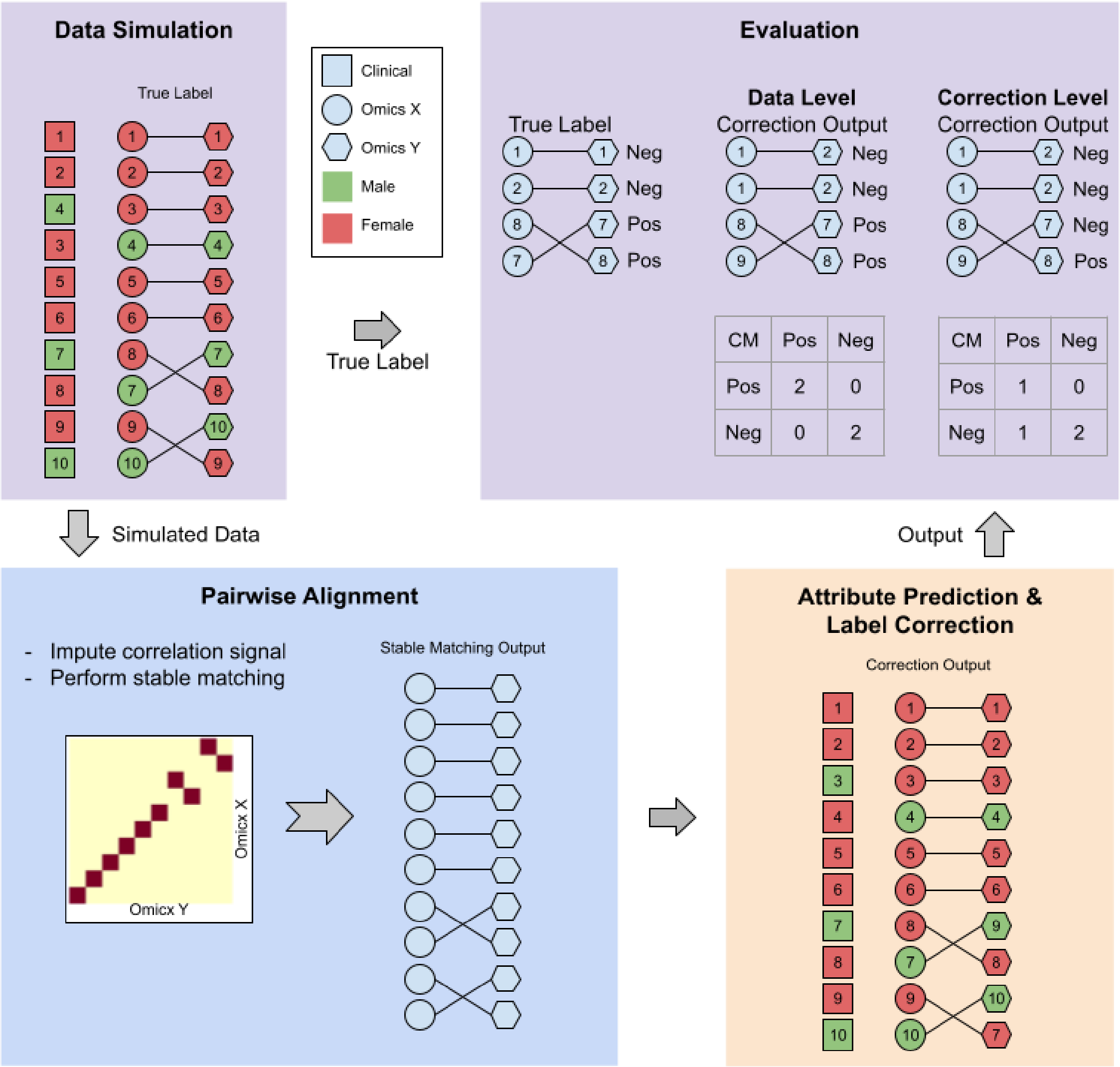

Swapping
1
2
1
2
1
2
Detect Swapping Pair
1
2
1
2
1
2
Checking annotated Label with Predicted Label
1
2
1
2
2
1
Correct the label
Duplication
23
5
5
Detect pairing with high matching score
0.223
3
5
5
8
0.697
Inspect the highest correlation each data has with others
3
3
5
8
8
8
The data with higher correlation is corrected
Shifting
5
6
7
8
9
Identify the shifting chain
Checking annotated Label with Predicted Label
7
8
9
Correct the label

Figure 5.2: The automated correction algorithm for each type of mislabeling error.

5.3 Simulation and Evaluation

These different algorithms (pairwise alignments, attribute prediction and label correction) were integrated together and form an automated pipeline, known as COSMO (COrrection of Sample by Multi Omics). To investigate the performance of COSMO in correcting individual labels, similar simulation experiments were conducted. A dataset was simulated with an error rate ranging from 0.05 to 0.28 and then the COSMO pipeline was used on the dataset. The process was repeated 50 times. The performance of COSMO was evaluated using F1 score in three levels: sample level, data level and correction level. Sample level F1 score is the same scores obtained in Chapter 4, which evaluate performance in detecting mislabeling of the patients' samples. Data level F1 score evaluates how accurately the pipeline identifies mislabeled omics type. In the data level, a mislabeled instance is treated as a true positive if and only if COSMO detects correctly which omics data get mislabeled. Correction level F1 score evaluates how accurate the pipeline corrects the label. In the correction level, the corrected label has to be the same with its true label in order to be treated as a true positive instance. Figure 5.4 represents the number of mislabeled samples with different types of error in each simulation. The error rate ranges from 0.5 to 0.28.



CM	Pos	Neg
Pos	2	0
Neg	0	2

CM	Pos	Neg
Pos	1	0
Neg	1	2

Figure 5.3: Simulation to evaluate the correction performance of COSMO. The simulation was repeated 50 times with various error rates.

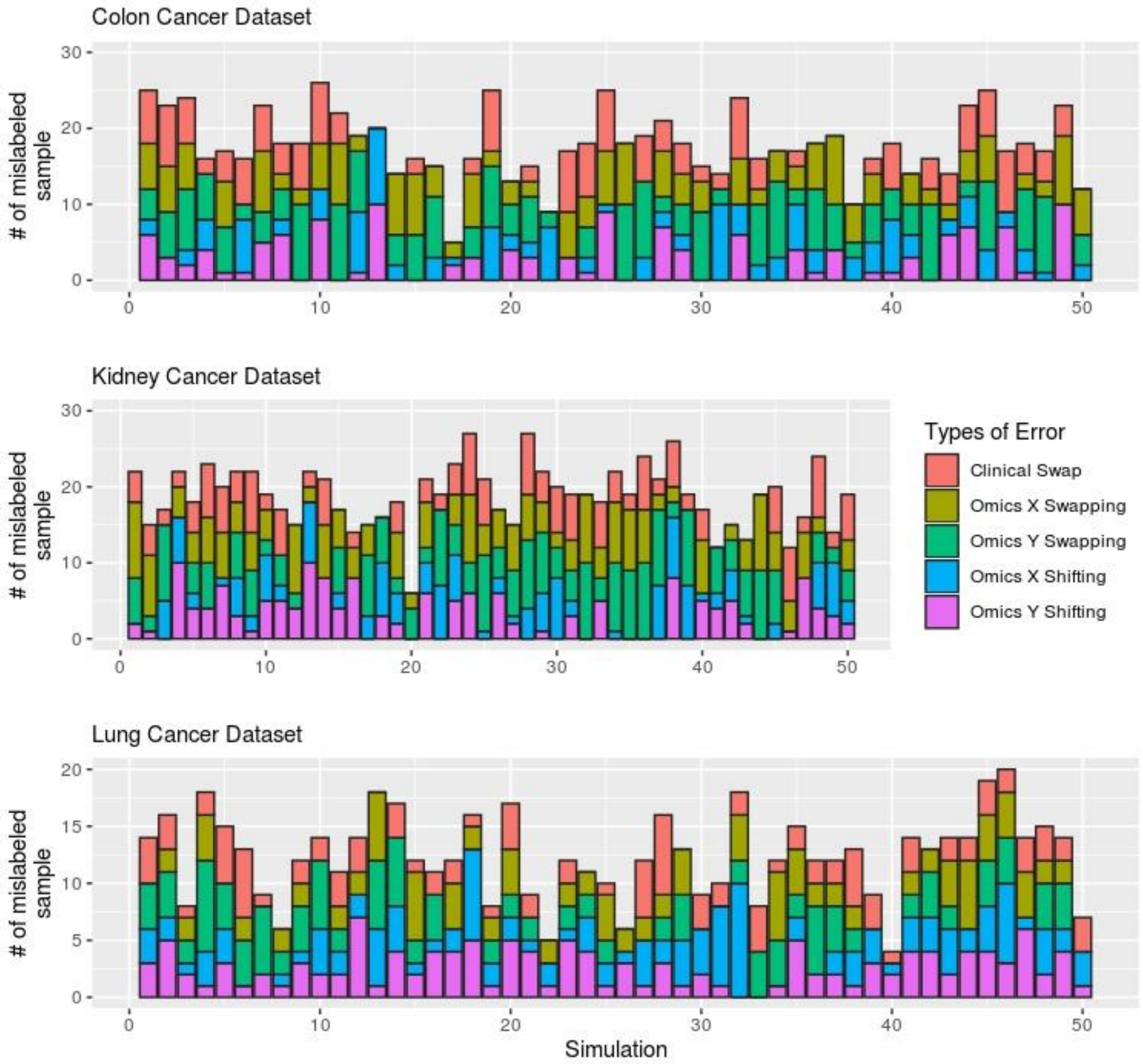


Figure 5.4: Number of mislabeled samples in each simulation across different datasets. The error rate varies from 0.05 to 0.28 in every simulation.

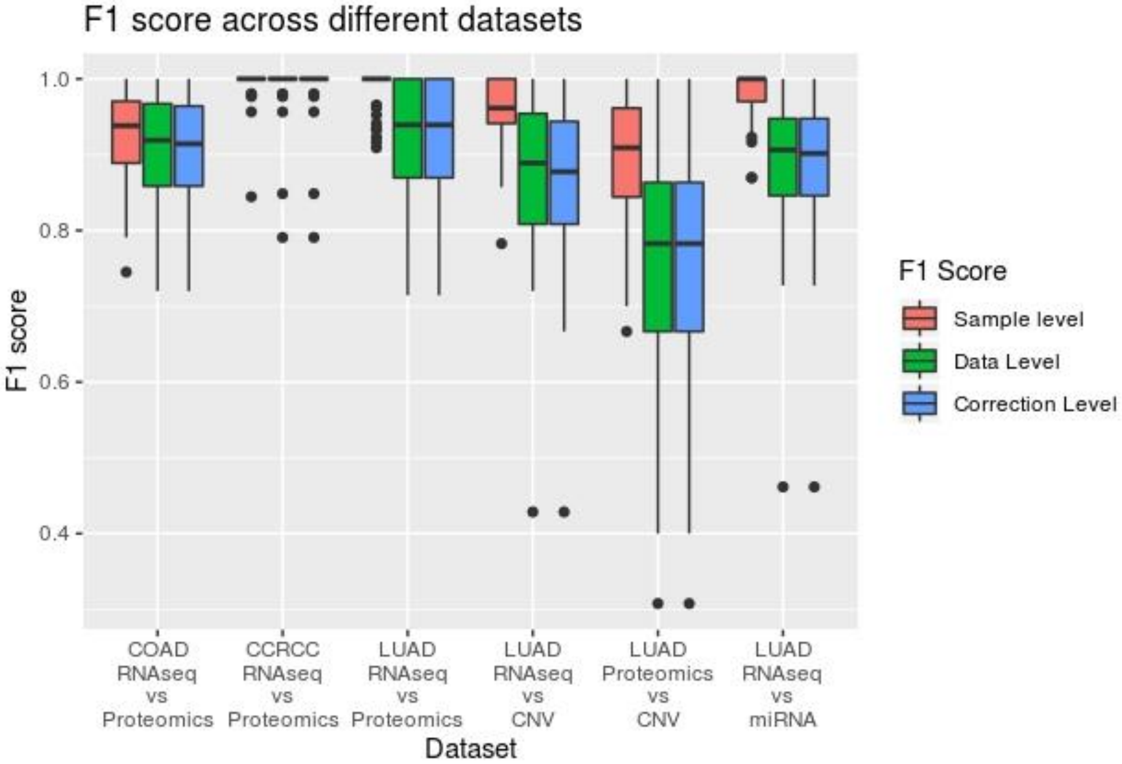


Figure 5.5: F1 scores of label correction across different pairs of omics data.

Dataset	Type of Omics	Type of Omics	Sample Level		Data Level		Correction Level	
			Mean	Median	Mean	Median	Mean	Median
COAD	RNAseq	Proteomic	0.9217	0.9381	0.9123	0.9189	0.9067	0.9143
CCRCC	RNAseq	Proteomic	0.9947	1.0000	0.9906	1.0000	0.9906	1.0000
LUAD	RNAseq	Proteomic	0.9880	1.0000	0.9252	0.9393	0.9252	0.9393
LUAD	RNAseq	CNV	0.9586	0.9615	0.8700	0.8889	0.8631	0.8775
LUAD	Proteomic	CNV	0.9001	0.9091	0.7591	0.7826	0.7591	0.7826
LUAD	RNAseq	miRNA	0.9836	1.0000	0.8899	0.9062	0.8891	0.9016

Table 5.1: Mean and Median of F1 scores in sample, data and correction level

Figure 5.5 shows that when aligning RNAseq and proteomics samples, COSMO achieved high average F1 scores in correcting mislabeled samples: 0.91, 0.99 and 0.92 in COAD, CCRCC and LUAD datasets in both data and correction level. When aligning RNAseq to CNV samples, COSMO achieved an average F1 score of 0.87 in data level and 0.86 in correction level respectively. While aligning RNAseq to miRNA samples, COSMO achieved an average F1 score of 0.89 in both levels. The results indicate that aligning RNAseq to any other omics data is accurate, making RNAseq data the most utilisable in detecting mislabeling. On the other hand, when aligning Proteomics to CNV data, COSMO achieved an average F1 score of 0.76.

To determine the robustness of the pipeline in label correction, the simulation experiments were repeated with different error rates, e = [0.1, 1.0]. The experiments are performed using the CCRCC dataset, as it has the highest F1 scores among all datasets. Figure 5.6 shows the F1 scores achieved against different error rates. The pipeline achieved a mean F1 score of 1.0 when error rate = 0.1. The mean F1 score decreased with the increasing error rate and correction level F1 score = 0.91 when error rate = 0.5. However, the pipeline is unable to perform label correction when error rate > 0.5, due to insufficient correctly labeled training data to train a classifier model. Without attribute prediction, the pipeline is unable to perform label correction, albeit still able to detect mislabeled samples (as shown in Chapter 4).

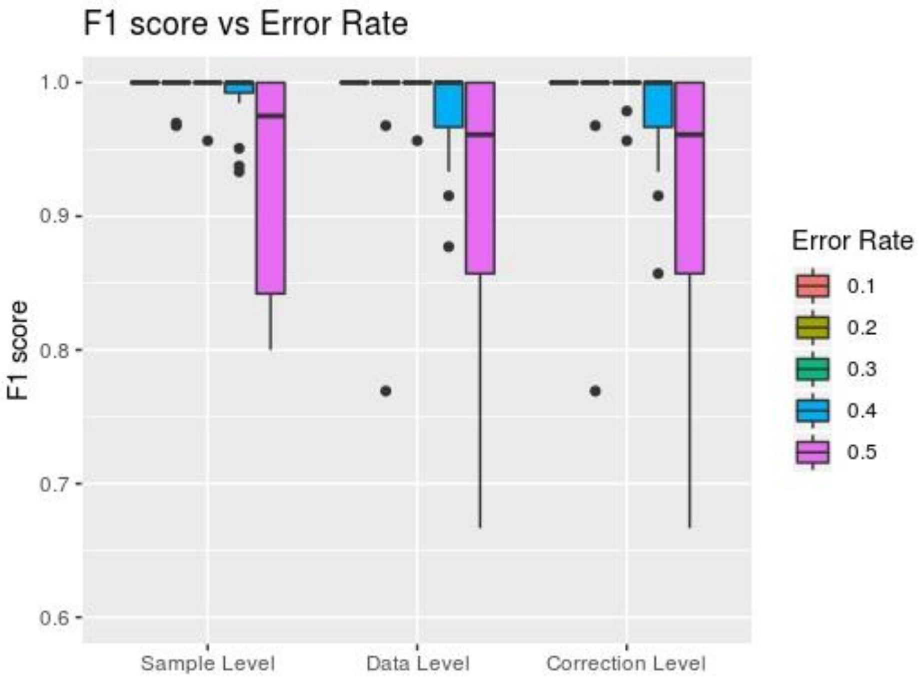


Figure 5.6: F1 scores against different error rates.

5.4 Real Case Study - Mouse Proteogenomic Dataset

The mouse proteogenomic dataset was collected from Chick et al. (2016). This dataset was derived from 192 mouse liver tissues and contains two types of omics data: RNAseq and proteomics. The RNAseq data contains 21321 gene features with no missing values. The proteomics data contains 8246 protein features: 1640 of them were removed due to missing rate > 50%, any remaining missing values were imputed via a random-forest based imputation, resulting in a total of 6606 protein features. The COSMO pipeline was used on the dataset and 20 samples were found to be mislabeled.

Among the 20 mislabeled samples, 18 were swapped (9 swapping pairs) and 2 were duplicated. The annotated label and predicted labels of those swapped samples were shown in Table 5.2. Four pairs were found to be proteomic swapping while the remaining five pairs were unknown due to same-sex sample swapping.

Upon further inspection, it was observed that all 20 mislabeled samples are from two different Tandem Mass Tag (TMT) batches, ten samples from Batch S14 and ten from Batch S15. TMT multiplexing is a proteomic quantification technique where several samples are tagged with unique isobaric tags, then are mixed and analyzed in a single liquid chromatography-mass spectrometry (LC-MS) experiment. The quantification of protein abundance and sample separation are then carried out in-silico. Every swapping case occurred between two samples from different batches. There is reasonable evidence to suggest that these two TMT batches got swapped, resulting in proteomic swapping in all these 20 samples.

TMT Batch	Sample	Annotated Label		Predicted Label RNAseq		Predicted Label Proteomic		Type of Swapped data
		sex	prob*	sex	prob*	sex	prob*	
S14	s_130FS	F	0	F	0.0039	F	0.0811	Unable to infer
S15	s_140FH	F	0	F	0.0147	F	0.0055	
S14	s_131FS	F	0	F	0.0072	F	0.0308	Unable to infer
S15	s_141FH	F	0	F	0.0089	F	0.0374	
S14	s_132FS	F	0	F	0.0043	F	0.4644	Unable to infer
S15	s_142FH	F	0	F	0.0129	F	0.0045	
S14	s_133FH	F	0	F	0.0163	F	0.1177	Unable to infer
S15	s_143FH	F	0	F	0.0103	F	0.0541	
S14	s_135MS	M	1	M	0.8622	M	0.5750	Proteomic
S15	s_146FS	F	0	F	0.0112	M	0.8641	
S14	s_136MS	M	1	M	0.9955	F	0.3709	Proteomic
S15	s_147FH	F	0	F	0.0565	M	0.9976	
S14	s_137MS	M	1	M	0.9894	M	0.9999	Unable to infer
S15	s_148MH	M	1	M	0.9948	M	0.9992	
S14	s_138MH	M	1	M	0.9779	F	0.0754	Proteomic
S15	s_149FH	F	0	F	0.0108	M	0.9844	
S14	s_139MS	M	1	M	0.9944	F	0.2745	Proteomic
S15	s_150FH	F	0	F	0.0167	M	0.9995	

Table 5.2: The annotated and predicted labels of 18 swapped samples. Every two consecutive samples, indicated by different colored cells, are swapped with each other. The prediction probability is the probability of being a male sample. The table is adapted from Yoo et al. (2021).

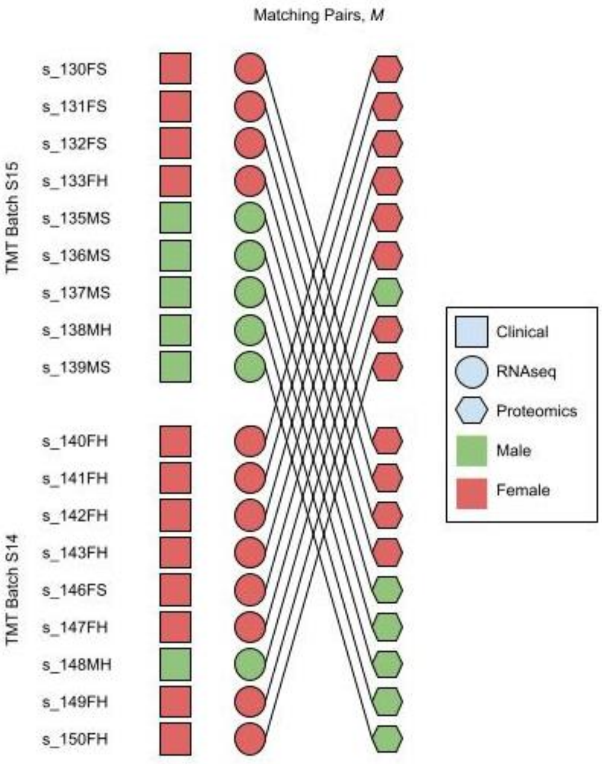
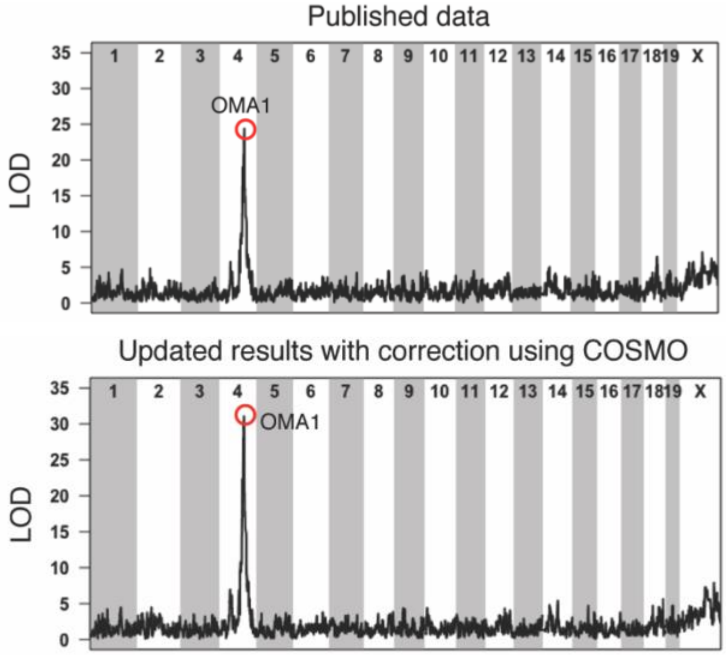


Figure 5.7: Four pairs of swapping were determined to be proteomics swapping. Though it is impossible to determine the source of error for another 5 pairs of swapping, the observation that every swapping occurred between samples from two different TMT batches suggests the Proteomics data get swapped during TMT multiplexing process.

The mislabeling rate in this dataset is 10.4% (20/192). To determine the impact of mislabeling in the analysis, the protein Quantitative Trait Loci (pQTL) analysis was rerun on the corrected data. In the published data, the most significant association is the OMA1 protein expression of OMA1 to a genetic marker in Chromosome 4 with a log odd ratio of 24. After correcting the data, the log odd ratio increased to 31, an increment by 1.3 fold (Figure 5.8).


Published data

Updated results with correction using COSMO

Figure 5.8: LOD score of OMA1 protein expression with a genetic marker in Chromosome 4 increased from 24 to 31 upon rerun the pQTL analysis on corrected data The figure was adapted from Yoo et al. (2021).

5.5 Discussion

The COSMO pipeline is useful in detecting the mislabeled samples and determining the source of error before correcting the labels. It was shown to have achieved high F1 scores in both data level and correction level, on three simulated cancer datasets. Mislabeling correction has the highest F1 score when comparing RNAseq data to Proteomics data, with an average score > 0.9 (both data and correction level) in all three datasets. It was observed that COSMO achieved the highest average F1 score of 0.99 in Kidney Cancer Dataset (CCRCC). This could be due to the higher number of protein features (CCRCC has 10127 protein features while COAD has 4105 and LUAD has 7556) which helps in extracting more reliable correlation signals. Mislabeling correction in RNAseq to CNV or to miRNA data also has a high average F1 score (> 0.86 in both data and correction level). This shows that RNAseq is the most utilisable omics data in detecting and correcting mislabeled samples. On the other hand, mislabeling corrections in Proteomics to CNV data achieved an average F1 score of 0.76. There is still room for improvements. Fortunately, this does not diminish the impact of COSMO as it still has a high detection capability (as shown in Chapter 4). Besides, RNAseq data is the most common omics data to be characterized in the research settings and most of the studies have RNAseq data which could be utilized in correcting Proteomics and CNV data.

To showcase the impact of COSMO, the pipeline was carried out on a real proteogenomic dataset to perform a quality check. It was found that 10.4% (20/192) of proteomics data get mislabeled. The pQTL analysis was rerun on corrected proteomics data and the most significant association (OMA1 protein expression with a genetic variant in Chromosome 4) has increased LOD score from 24 to 31, showing that a small mislabeling rate of 10.4% can have an impact of reducing the significance by 23% (7 / 31) on the LOD score.

6 MISLABELING CORRECTION IN MULTI OMICS DATA

COSMO has achieved good performance in detecting and correcting mislabeling, but its application is limited to datasets with exactly two omics data. Theoretically, a study with more than two omics data could perform the same pipeline iteratively for different pairs of omics data, but each application only utilizes information from the two omics data being inspected and do not approach the task in a holistic manner. The advantage of having more types of omics data is not fully exploitable since more dimensions of omics data is hypothesized to have more information in mislabeling handling. Besides, each application performs predictions for two omics data and it is redundant in consecutive applications where the same omics data were corrected. The redundancy decreases the efficiency of the quality control check and the reduction is even higher if a dataset has more types of omics data. In Chapter 6, an algorithm was proposed to integrate all omics data to detect and correct individual mislabeling in a dataset. The performance of the proposed algorithm was investigated, along with its application in real life datasets.

6.1 Network Topology Realignment

Considering a patient’s tissue sample was used to characterize three types of omics data (RNAseq, proteomics and CNV), these three different data instances should have the highest correlation signal with each other. In other words, three pairwise alignments will be performed: RNAseq to proteomics, RNAseq to CNV, and proteomics to CNV. In each alignment, the data instances from the same patient should have the highest correlation signal and are paired with each other.

Thus, the mislabeling detection and correction tasks could be approached as a network topology realignment task. An algorithm was proposed to integrate all pairwise alignments to perform the realignment. In the network, the set of vertices consists of all data instances of every omics data and the edge represents the matching pairs in each pairwise alignments. Data instances from the same patient should have the highest correlation signal with each other, and thus they should be connected with only each other, forming a tightly connected cluster on the network. These data instances are considered as correctly labeled whereas those that do not pair with itself are considered mislabeled. Then, those considered mislabeled will have their correlation to other omics inspected to perform label correction.

During label correction, the priority of the mislabeled instances were determined. Priority indicates the number of mismatches in each pairwise alignment. The instance with the highest priority will have its label get corrected first. To determine the correct label of the instance, its correlation to other instances from different omics was inspected for every patient and the one with the highest correlation will be the corrected label. As the instance gets corrected, it forms a new connected cluster. All the instances within the cluster will have their priority updated. The label correction process repeats until all the instances have their label corrected or their priority becomes zero.

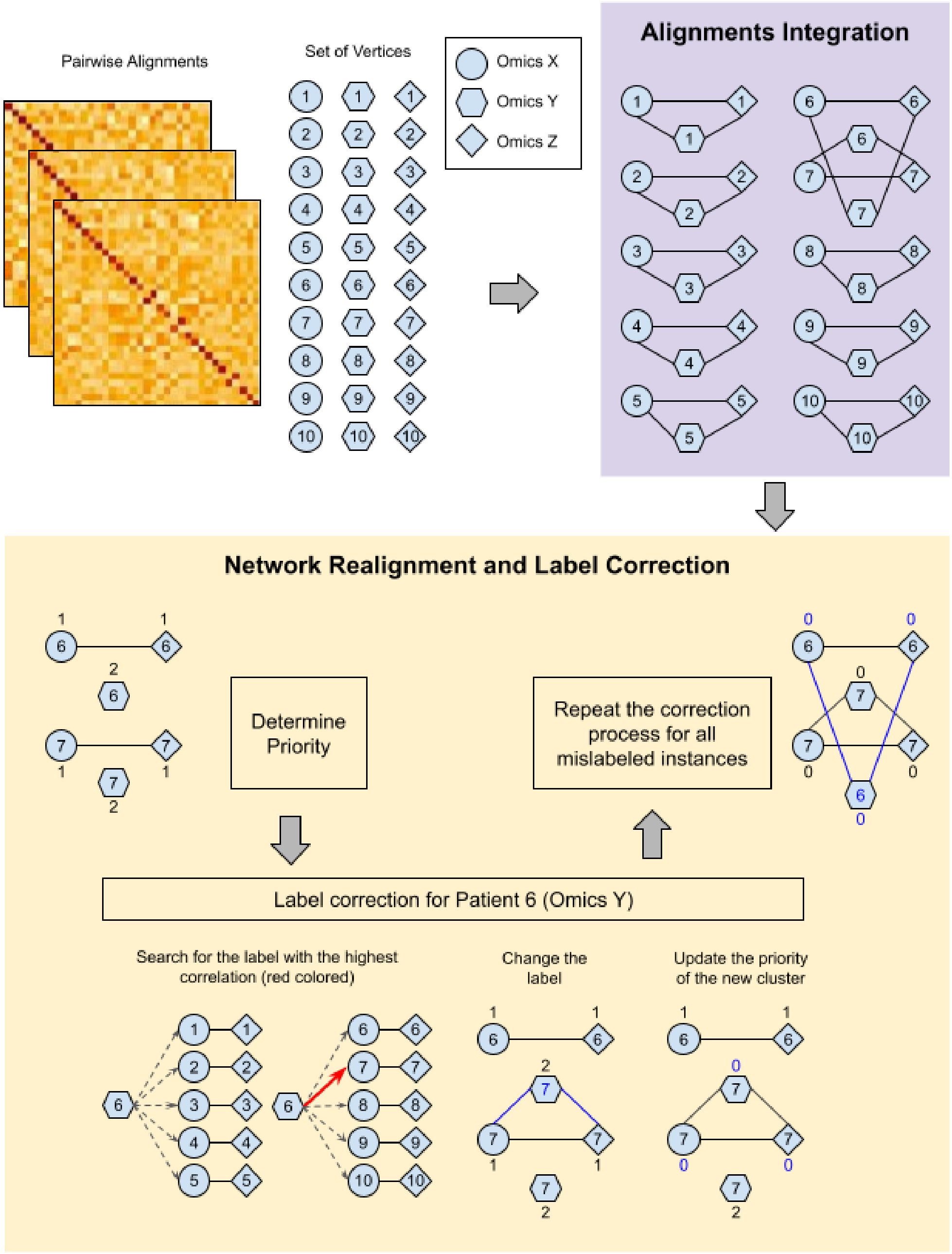

Pairwise Alignments
Set of Vertices
Omics X
Omics Y
Omics Z
Alignments Integration
Network Realignment and Label Correction
Determine Priority
Repeat the correction process for all mislabeled instances
Label correction for Patient 6 (Omics Y)
Search for the label with the highest correlation (red colored)
Change the label
Update the priority of the new cluster

Figure 6.1: Network Topology Realignment for correcting the labels for datasets with more than two types of omics data. Mismatched instances will get their labels corrected. The correction process keeps iterating until all mismatched instances get corrected or become zero priority.

6.2 Simulation and Evaluation

A mislabeled dataset was simulated from LUAD dataset with four omics data: RNAseq, proteomics, CNV and miRNA data. A total of 100 patients were randomly selected and mislabeling errors were artificially introduced to the dataset. Pairwise alignments were performed on the simulated dataset and the alignments output were used to handle mislabeling via network topology realignment. The performance was evaluated using F1 scores in three levels: sample level (if a sample has mislabeled data), data level (if a data instance is mislabeled), and correction level (if a mislabeled instance is corrected to its true label). The process was repeated 10 times for each error rate where error rate = [0.1, 1.0].

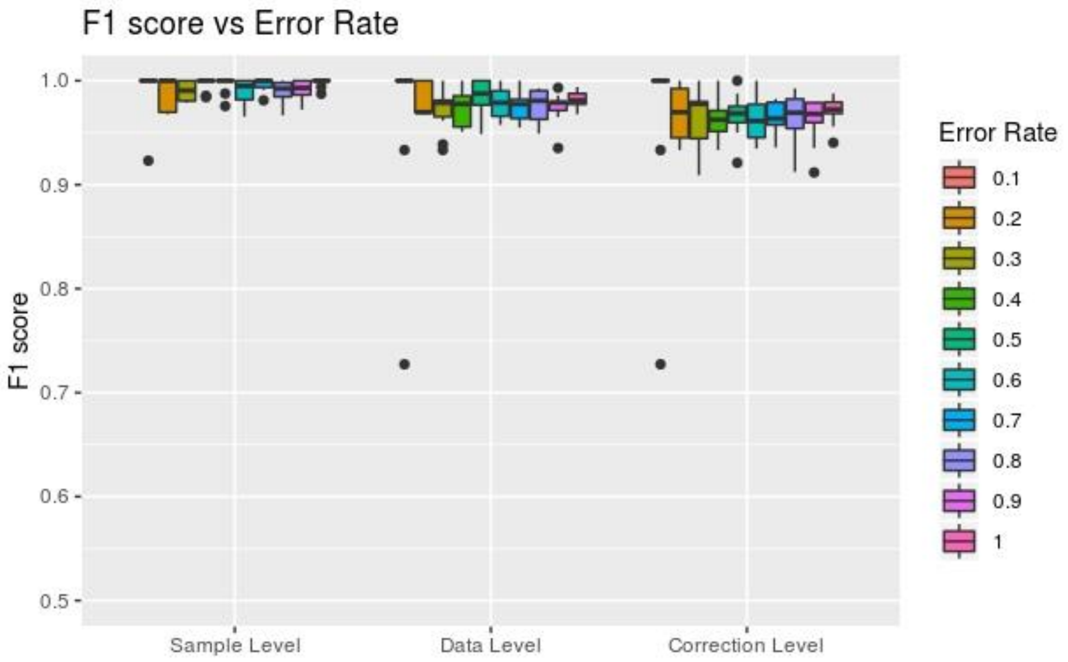


Figure 6.2: F1 scores of network topology realignment in multi omics data across different error rates.

Error Rate	Sample Level		Data Level		Correction Level	
	Mean	Median	Mean	Median	Mean	Median
0.1	0.9846	1.0000	0.9458	0.9706	0.9458	0.9706
0.2	0.9937	1.0000	0.9709	0.9697	0.9603	0.9687
0.3	0.9921	1.0000	0.9554	0.9583	0.9421	0.9545
0.4	0.9986	1.0000	0.9669	0.9677	0.9605	0.9677
0.5	0.9938	1.0000	0.9656	0.9740	0.9522	0.9538
0.6	0.9916	0.9899	0.9658	0.9684	0.9488	0.9519
0.7	0.9953	1.0000	0.9515	0.9557	0.9369	0.9423
0.8	0.9937	0.9963	0.9545	0.9593	0.9388	0.9447
0.9	0.9959	0.9966	0.9578	0.9603	0.9364	0.9382
1.0	0.9969	0.9971	0.9638	0.9697	0.9443	0.9425

Table 6.1: Mean and Median F1 scores of label correction across different error rates.

Figure 6.2 shows that the network realignment algorithm achieved high F1 scores (> 0.99 in patient level, > 0.95 in data level and > 0.94 in correction level) across different error rates. The algorithm is robust against error rates even at the correction level. This emphasizes the advantage of having more than two omics data, the misalignment of one data could be corrected by other pairwise alignments. It should be noted that the error rate indicates the proportion of samples with mislabeled data. In other words, a dataset with an error rate of 1.0 indicates that every sample has one mislabeled data. The label correction is still feasible in such high error rate as the mislabeled data could be realign to other correctly labeled omics data.

6.3 Real Case Study

6.3.1 TCGA Breast Cancer Dataset (BRCA)

TCGA is the largest public multi omics repository to date. Breast Invasive Carcinoma dataset is the largest cancer dataset to date. It consists of data of 521 patients with all three types of omics data collected: RNAseq, microarray and CNV. The omics quantification pipelines were described in detail on the website[9]. The omics data contains 20501, 17274, and 25187 gene features respectively with no missing values. Three pairwise alignments were performed and network realignment identified 16 mislabeled samples. These 16 samples were swapped in microarray data and the samples were listed in Table 6.2.

Breast cancer dataset is highly gender imbalanced in that most of the samples belong to female patients. Only 0.1% (6 / 521) of the samples belong to male patients. Table 6.2 shows that one mislabeled sample belongs to male patient. To investigate the impact of the mislabeling, the differential expressed gene (DEG) analysis was run between male and female patients. T-test was performed on each gene feature between two groups and the significance of the gene was adjusted by Benjamini-Hochberg procedure. The analysis was run twice, first on the mislabeled data and then corrected data. The number of

[9] https://docs.gdc.cancer.gov/Data/Introduction/

DEGs are shown in Figure 6.3. In the first run, 16 genes were found to be differentially expressed while in the second run, the number of DEG increased to 59, an increment of 3.7 fold. Among these two sets of DEGs, 13 of them overlapped which means the mislabeling prevented the discovery of 46 true DEGs. Three genes, previously thought to be significant, are in fact not significant (Figure 6.4).

First sample in a pair		Second sample in a pair	
Label	Sex	Label	Sex
TCGA.BH.A0BA	Female	TCGA.BH.A0DS	Female
TCGA.BH.A18K	Female	TCGA.BH.A18T	Female
TCGA.BH.A0BS	Female	TCGA.BH.A0BT	Female
TCGA.AR.A1AW	Female	TCGA.AR.A1AV	Male
TCGA.BH.A0H3	Female	TCGA.BH.A0HA	Female
TCGA.E2.A1B5	Female	TCGA.E2.A1B6	Female
TCGA.AR.A1AN	Female	TCGA.AR.A1AL	Female
TCGA.BH.A0EI	Female	TCGA.A1.A0SD	Female

Table 6.2: Eight swapping pairs of 16 mislabeled microarray data. Each row represents each swapping pair.

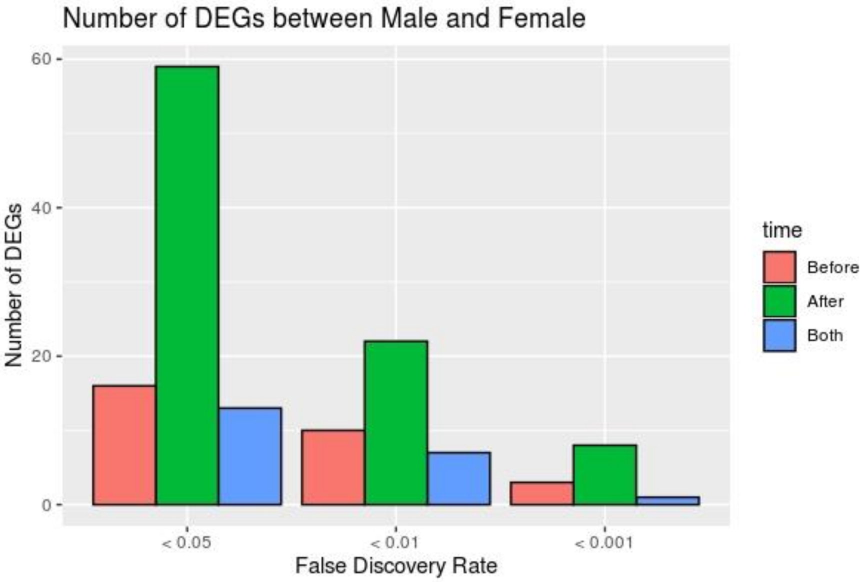


Figure 6.3: Number of DEGs before and after correction across different thresholds.

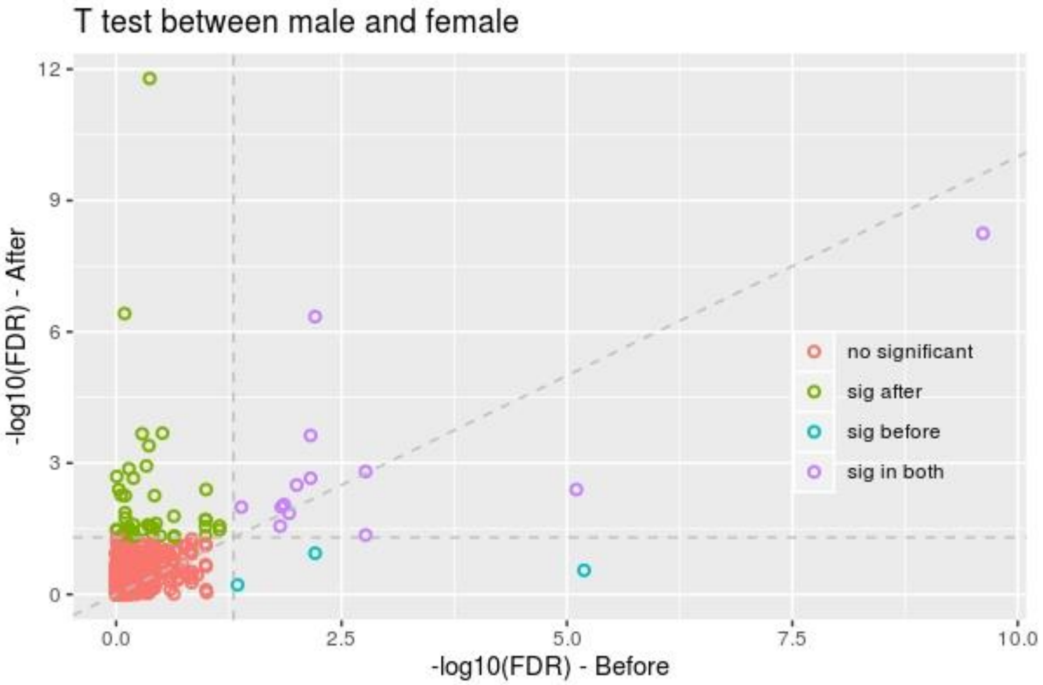


Figure 6.4: Change of false discovery rate of genes before and after the correction. Horizontal and vertical grey lines indicate the FDR cutoff of 0.05.

6.3.2 Lymphoblastoid cell lines (LCLs) dataset

Battle et al. (2015) conducted a study to determine the association of genetic markers with expression quantitative trait loci (eQTLs), ribosome occupancy (rQTLs), or protein abundance (pQTLs). Three types of omics data were generated from Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) derived from 60 human individuals. RNAseq and Riboseq were quantified based on Fragments Per Kilobase of transcript per Million mapped reads (FPKM) as described in the study. Protein abundance was measured using a SILAC internal standard sample (Ong et al., 2002) and quantitative protein mass spectrometry. RNAseq data contains 16614 gene features while the riboseq data contains 15059 genes with no missing values. The proteomics data contains 4381 proteins and the missing values were imputed by random forest based imputation.

Battle et al. (2015) observed that many QTLs exhibit shared effects across mRNA, ribosome occupancy (riboseq) and protein, indicating that riboseq has reasonable correlation to RNAseq and proteomics data. Three pairwise alignments were performed and the distribution of gene correlation across different pairs of omics data were inspected. Figure 6.5 shows that the Pearson Correlation follows normal distribution in every pair of omics data. Riboseq is a high throughput method based on deep sequencing of ribosome-protected mRNA fragments, providing an estimate of protein translation efficiency. Thus, riboseq has a higher mean correlation to RNAseq and to Proteomics, compared to the mean correlation of RNAseq to proteomics data.

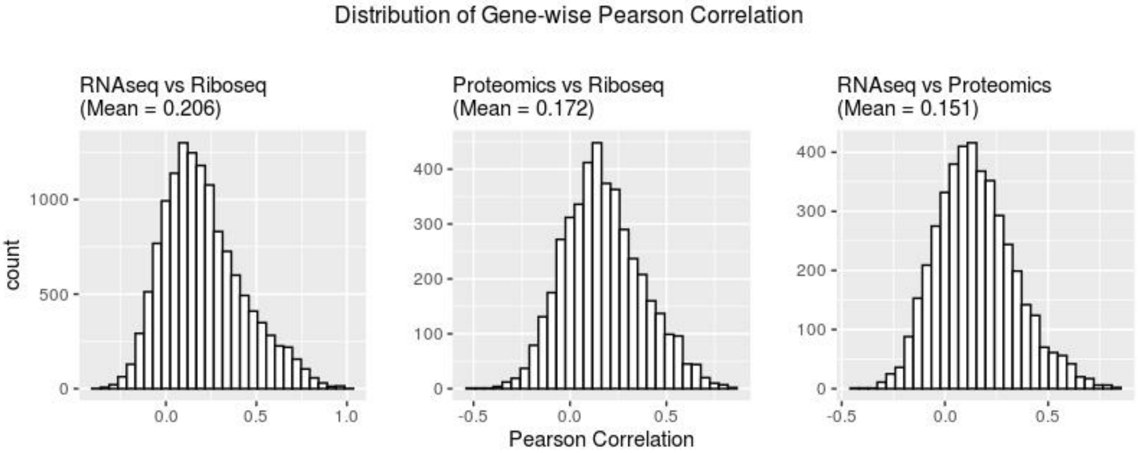


Figure 6.5: Distribution of gene-wise Pearson Correlation across different pairs of omics data.

The alignment outputs were feedforward to the network realignment algorithm. Three samples were found mislabeled. Two RNAseq data were found to be swapped, while one proteomics data was found to be duplicated. The mislabeling rate of the dataset is 5% (3 / 60).

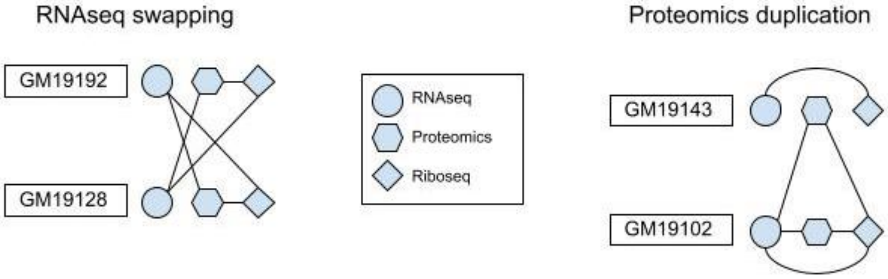
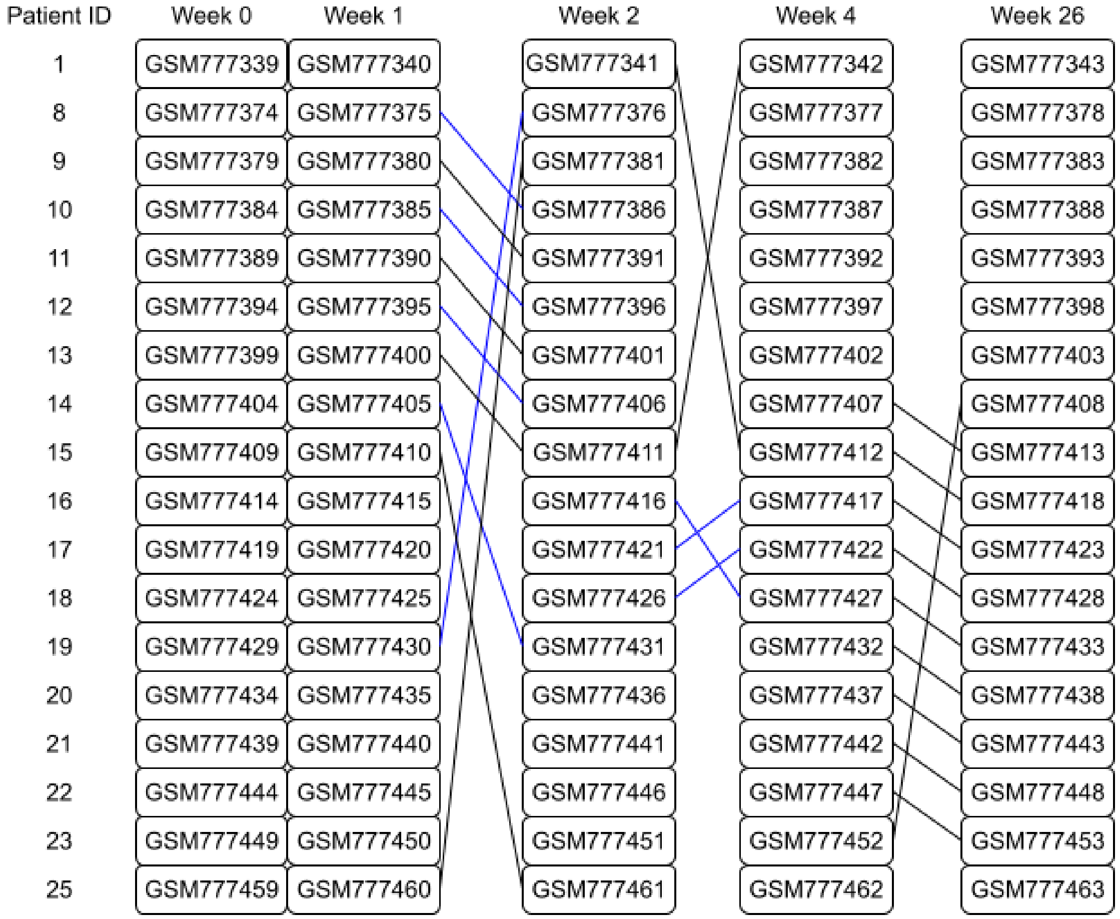


Figure 6.6: Three Mislabeled samples found in LCLs dataset.

6.3.3 Tuberculosis Patients Blood Gene Expression

Cliff et al. (2013) conducted a study on 27 tuberculosis patients with the aim to differentiate blood gene expression with respect to the treatment response. The patients were given conventional therapy (2HRZE/4HR) for 6 months: isoniazid, rifampin, pyrazinamide and ethambutol for 2 months (2HRZE) followed by isoniazid plus rifampin for 4 months (4HR). During the treatment duration, blood samples were collected in five different timepoints: prior to starting standard therapy and after 1, 2, 4, and 26 weeks of successful treatment. In total, 135 blood samples were collected (27 patients × 5 timepoints). The gene expression of blood samples was characterized using the microarray platform A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0, measuring expression of 21319 genes in every sample.

There is only one type of omics data in this dataset (transcriptomics) but the patients have their blood samples collected five times in five different timepoints. Ten pairwise alignments were performed across all different pairs of timepoints and the alignment outputs were used for network realignment. A total of 25 samples from 17 patients were found to be mislabeled, as shown in Figure 6.7. No sample was mislabeled at the time of diagnosis and one week after treatment. In the second week, 10 samples were mislabeled in two shifting events (each shifting involved 5 samples). In the fourth week, 5 samples were mislabeled: two of them swapped while three got shifted. At the last timepoint 26th week, 10 samples got shifted in one shifting event.



Patient ID	Week 0	Week 1	Week 2	Week 4	Week 26
1	GSM777339	GSM777340	GSM777341	GSM777342	GSM777343
8	GSM777374	GSM777375	GSM777376	GSM777377	GSM777378
9	GSM777379	GSM777380	GSM777381	GSM777382	GSM777383
10	GSM777384	GSM777385	GSM777386	GSM777387	GSM777388
11	GSM777389	GSM777390	GSM777391	GSM777392	GSM777393
12	GSM777394	GSM777395	GSM777396	GSM777397	GSM777398
13	GSM777399	GSM777400	GSM777401	GSM777402	GSM777403
14	GSM777404	GSM777405	GSM777406	GSM777407	GSM777408
15	GSM777409	GSM777410	GSM777411	GSM777412	GSM777413
16	GSM777414	GSM777415	GSM777416	GSM777417	GSM777418
17	GSM777419	GSM777420	GSM777421	GSM777422	GSM777423
18	GSM777424	GSM777425	GSM777426	GSM777427	GSM777428
19	GSM777429	GSM777430	GSM777431	GSM777432	GSM777433
20	GSM777434	GSM777435	GSM777436	GSM777437	GSM777438
21	GSM777439	GSM777440	GSM777441	GSM777442	GSM777443
22	GSM777444	GSM777445	GSM777446	GSM777447	GSM777448
23	GSM777449	GSM777450	GSM777451	GSM777452	GSM777453
25	GSM777459	GSM777460	GSM777461	GSM777462	GSM777463

Figure 6.7: Mislabeling of 25 samples from 17 patients. Patients whose samples are not shown here are correctly labeled. There is no mismatched sample in Week 0 and Week 1. In week 2, there are two shifting events, indicated by different colors of edges. In week 4, the shifting events are indicated by blue-colored edges. Only mismatched edges are shown here.

The paired t-test was conducted between Week 4 and Week 26 after label correction. The number of genes found to be differentially expressed is 3844 before correction and 4019 after label correction. Among these two sets of DEGs, 3709 genes overlapped. In other words, the mislabeling prevented the discovery of 310 genes, 7.71%

(310 / 4019) of the total number of genes. There are 135 previously identified DEGs that are in fact not significant (Figure 6.9).

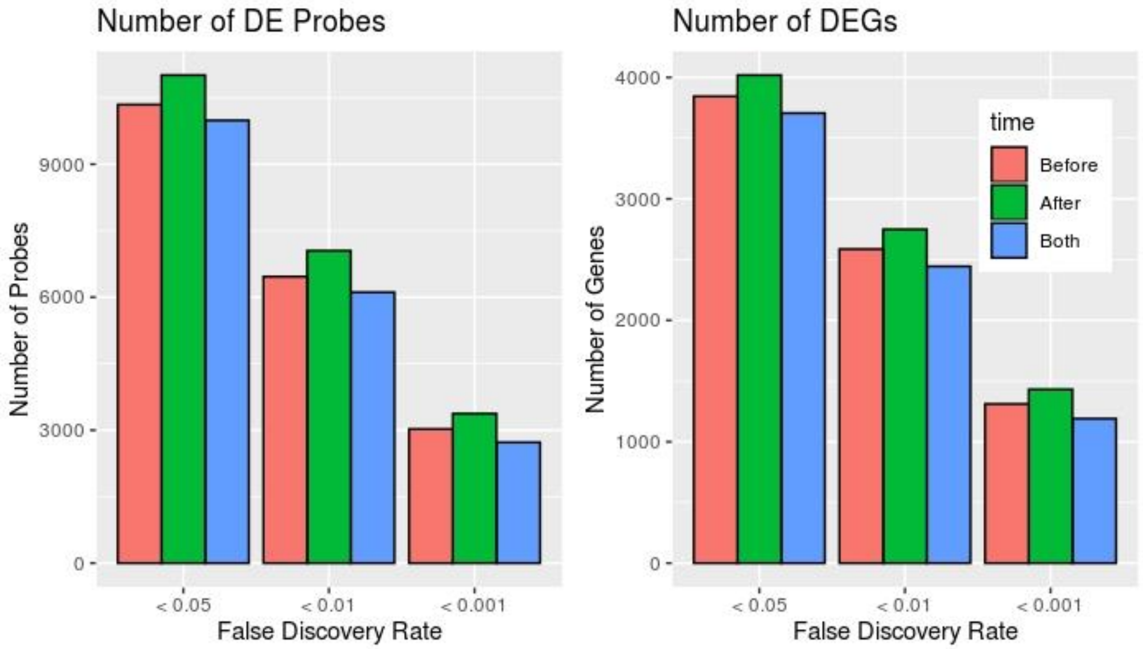


Figure 6.8: Number of differentially expressed probes (left) and genes (right) between Week 4 and Week 26 before and after the label correction.

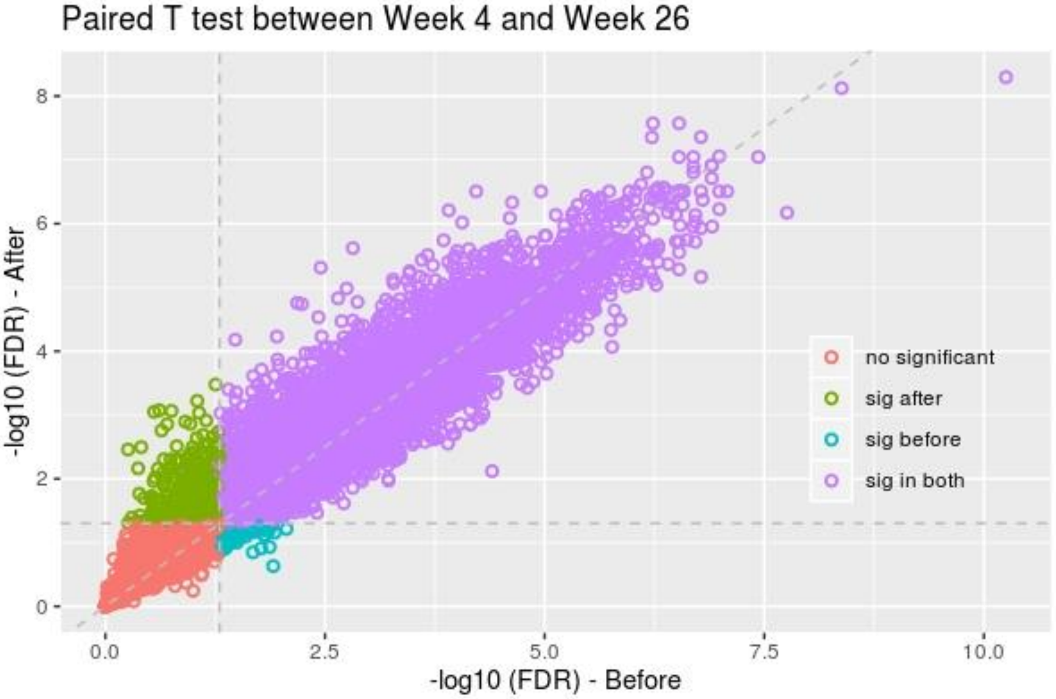


Figure 6.9: Change of false discovery rate of probes before and after the correction. Horizontal and vertical grey lines indicate the FDR cutoff of 0.05.

6.4 Prevalence

The mislabeling pipeline was applied on other public datasets to detect any mislabeling and determine the mislabeling rate overall. Cancer Cell Line Encyclopedia (CCLE) is a multi omics repository of several human cancer cell lines (Ghandi et al., 2019). To date, the repository contains data of 1457 cell lines. Three types of omics data were collected from the website[10]: RNAseq, proteomics and CNV. Cell lines that do not have all three omics data were filtered, leaving 371 cell lines at the end. Three pairwise alignments were

[10] https://portals.broadinstitute.org/ccle

performed on the dataset followed by network realignment. All cell lines aligned perfectly with each other and no mislabeling was found.

TCGA is the largest public multi omics repository to date, consisting of 38 different cancers. Nine cancer datasets having all four types of omics data were collected (RNAseq, microarray, CNV and miRNA) and the mislabeling detection algorithm was applied on these datasets. Table 6.3 shows the summary statistics of mislabeled samples. The mislabeled samples were found in 5 (55.55%) out of 9 cancer datasets. In total, there are a total of 1259 subjects and 44 (3.49%) of them whose data has been mislabeled. The mislabeling rates vary across datasets, ranging from 0% to 28.13%.

Cancer Dataset	Number of Subjects	Subjects with mislabeled data	Omics Assay	Sample Size	Mislabeled Assay	Rate (%)
BRCA	521	18 (3.45%)	RNAseq Microarray CNV miRNA **Total**	521 521 521 312 **1875**	0 16 0 2 **18**	0 3.07 0 0.64 **0.96**
COAD	135	8 (5.93%)	RNAseq Microarray CNV miRNA **Total**	135 135 135 121 **526**	0 5 0 3 **8**	0 3.70 0 2.48 **1.52**
GBM	19	0 (0%)	RNAseq Microarray CNV **Total**	19 19 18 **56**	0 0 0 **0**	0 0 0 **0**
KIRC	71	4 (5.63%)	RNAseq Microarray CNV	71 71 69	0 3 0	0 4.23 0

			miRNA	62	1	1.61
			Total	**273**	**4**	**1.47**
KIRP	16	0 (0%)	RNAseq	16	0	0
			Microarray	16	0	0
			CNV	16	0	0
			miRNA	16	0	0
			Total	**64**	**0**	**0**
LGG	27	0 (0%)	RNAseq	27	0	0
			Microarray	27	0	0
			CNV	27	0	0
			miRNA	27	0	0
			Total	**108**	**0**	**0**
LUAD	32	9 (28.13%)	RNAseq	32	0	0
			Microarray	32	9	28.13
			CNV	32	0	0
			miRNA	32	0	0
			Total	**128**	**9**	**7.03**
LUSC	151	0 (0%)	RNAseq	151	0	0
			Microarray	151	0	0
			CNV	151	0	0
			miRNA	127	0	0
			Total	**580**	**0**	**0**
OV	287	5 (1.74%)	RNAseq	287	0	0
			Microarray	287	0	0
			CNV	284	0	0
			miRNA	280	5	1.79
			Total	**1138**	**5**	**0.44**
Total	1259	44 (3.49%)		4748	44	0.93

Table 6.3: Summary of mislabeled samples found in TCGA datasets.

Expression Atlas is an archive storing gene expression data from high-throughput experiments housed by EBI (European Bioinformatics Institute). Expression Atlas stores and displays gene expression data across species and biological conditions, which enables the user to retrieve desirable datasets. The mislabeling detection algorithm is applicable to datasets in which every subject contributes at least 2 samples. The datasets of *homo sapiens* from differential experiments were collected where the experimental factor is time or treatment. Datasets were filtered based on these criterias: 1) contributed by too few subjects (< 8 subjects and < 16 assays), 2) the assays are generated by the same cell line and no heterogeneity in different samples. A total of 48 datasets were collected, consisting of 6900 assays contributed by 1993 subjects. The application of the proposed approach revealed eight datasets (16.67%) to have mislabeled data (as listed in Table 6.4 and appendix A). In overall, 36 (1.81%) subjects have their data mislabeled, affecting a total of 44 (0.64%) assays.

	Number of Subjects	Subjects with Mislabeled Data	Number of Assays	Mislabeled Assays
E-MTAB-6558	107	3 (2.80%)	288	3 (1.04%)
E-GEOD-19519	112	1 (0.89%)	224	1 (0.45%)
E-MTAB-7032	61	3 (4.92%)	158	3 (1.90%)
E-GEOD-41168	42	3 (7.14%)	140	3 (2.14%)
E-GEOD-31348	27	17 (62.96%)	135	25 (18.52%)
E-GEOD-58558	19	2 (10.53%)	109	2 (1.83%)
E-GEOD-23597	42	1 (2.38%)	107	1 (0.93%)
E-TABM-1138	142	6 (4.23%)	284	6 (2.11%)

Table 6.4: Datasets collected from Expression Atlas which contain mislabeled data. Details of other datasets could be found in Appendix A.

To determine the overall prevalence of mislabeling in public omics datasets, the summary of the mislabeled samples was compiled and inspected in respect to studies, subjects and assays. The study refers to a project or an experiment which generates the dataset. The subject refers to the organism where the biological sample is collected from. In a multi-omic study, one subject contributes one biological sample; but in a multi-timepoint study, one subject could contribute multiple biological samples. Hence, the number of subjects instead of samples is compiled in this table. The assay refers to any omics data generated in the study. In a multi-omic study, one sample is sequenced for several omics types but in a multi-timepoint study, one sample is only assayed for one omics data. The number of assay is the number of data generated regardless of the types of omics.

Table 6.5 shows that 25% of datasets contain mislabeled data. These datasets consist of omics data from 3875 subjects and 105 (2.71%) subjects' data have been mislabeled. Looking further into the performed assays, 113 (0.85%) out of 13325 assays were mislabeled. Given that most of the multi-omic studies combine omics data to perform integrative analysis, it is more practical to look into the mislabeling rate in subject level than assay level. Although there is only less than 1% of mislabeled assays, it results in a much higher mislabeling rate in subject level, highlighting again the importance of performing quality check before data analysis.

	Number of Datasets/ Studies	Studies Containing Mislabeled Samples	Number of Subjects	Number of Subjects with mislabeled data	Number of Assays	Number of Mislabeled Assay
Chick et al. (2016)	1	1	192	20	384	20
TCGA	9	5	1259	44	4748	44
Battle et al. (2016)	1	1	60	5	180	5
CCLE	1	0	371	0	1113	0
Expression Atlas	48	8	1993	36	6900	44
Total	**60**	**15 (25%)**	**3875**	**105 (2.71%)**	**13325**	**113 (0.85%)**

Table 6.5: Overall mislabeling rate compiled from various sources of datasets.

7 CONCLUSION AND FUTURE WORK

Multi omics study is getting more common in recent years. It characterizes several types of omics data and takes a system biology approach to gain insight in understanding biological processes. The scale of the study is getting larger and more omics data are generated, contributed by the collaborative efforts of researchers. The large scale of the study does not come by without any consequences. Sample mislabeling is a prevalent problem in multi omics studies and has led to unwanted consequences: irreproducibility of the result, unnecessary research effort and cost, and the discovery of false claims.

While the large scale of the multi omics study poses a risk of sample mislabeling, the multi dimension of omics data generated in the study presents an opportunity to perform quality check, making sure the data is attributed to the correct label before doing any data analysis. In this thesis, the quality check is approached as alignment tasks. The omics data from the same patient that are aligned together are considered correctly labeled and vice versa. Chapter 4 shows that every individual contains a unique signal in the omics data that is useful in the sample alignment. A method was proposed to extract correlation / coexpression signals and the signals are shown to be reliable in performing sample alignment. The method was able to achieve F1 scores of at least 0.95 in detecting mislabeling and is robust against error rate.

The pairwise alignments outputs were further inspected to correct the individual labels of the samples in a study where only two types of omics data are available. A pipeline was proposed to integrate predicted sex genotype in the label correction task. Chapter 5 shows that the pipeline was able to achieve F1 scores of at least 0.88 in correcting the labels when comparing RNAseq data to any other types of omics data. The utilization of predicted genotype enables the label correction task and the limitation lies in this component as well. To accurately correct the label, the prediction has to be accurate and to be able to correct the label, the mislabeled samples should have the opposite class label in the first place to begin with. The pipeline was applied on a real dataset (Battle et al., 2015). Nine pairs of samples were found to be swapped, however, the algorithm was unable to determine the source of error for 5 pairs as the swapping occurred between same sex sample. Fortunately, manual inspection revealed the swapping occurred between two TMT batches during proteomic multiplexing measurements, suggesting the source of error is on proteomics data. Though only sex attribute was utilized in the work here, theoretically, any other attribute that could be accurately predicted from omics data should enable the label correction task as well.

For datasets with more than 2 types of omics data, an algorithm was proposed to realign the data. The algorithm was able to achieve F1 scores of at least 0.94 in correcting individual labels. Due to the presence of at least 3 types of omics data, more pairwise alignments are performed and one mislabeled data could be realigned with another omics data without the attribute prediction. This mitigates the limitation of the previous pipeline and showcases the advantage of having more dimensions of omics data. Besides, more dimensions of omics data enable the network realignment algorithm to be robust against error rate.

Several datasets were collected from public repositories and the correction pipelines were performed to detect the mislabeled samples. Overall, 2.71% (105 / 3875) of the subjects are found to have mislabeled data. Though most of the datasets were free of any mislabeling, one dataset was observed to have a mislabeling rate as high as 28.13%. This showed the significance of performing quality checks in multi omics studies. An automated correction algorithm was developed to detect and correct the mislabeled samples to the individual level. The omics data inspected in this work are transcriptomics, proteomics, CNV and miRNA. One of the future directions includes investigating the quality check pipeline for other omics data such as genomic, epigenomic, metabolomic, phosphoproteomic and others.

REFERENCES

Astion, M. L., Shojania, K. G., Hamill, T. R., Kim, S., & Ng, V. L. (2003). Classifying laboratory incident reports to identify problems that jeopardize patient safety. *American Journal of Clinical Pathology*, *120*(1), 18–26.

Baskerville, S., & Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* , *11*(3), 241–247.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., & Gilad, Y. (2015). Impact of regulatory variation from RNA to protein. In *Science* (Vol. 347, Issue 6222, pp. 664–667). https://doi.org/10.1126/science.1260793

Broman, K. W., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Sen, Ś., & Attie, A. D. (2015). Identification and Correction of Sample Mix-Ups in Expression Genetic Data: A Case Study. In *G3 Genes|Genomes|Genetics* (Vol. 5, Issue 10, pp. 2177–2186). https://doi.org/10.1534/g3.115.019778

Buyske, S., Yang, G., Matise, T. C., & Gordon, D. (2009). When a Case Is Not a Case: Effects of Phenotype Misclassification on Power and Sample Size Requirements for the Transmission Disequilibrium Test with Affected Child Trios. In *Human Heredity* (Vol. 67, Issue 4, pp. 287–292). https://doi.org/10.1159/000194981

Chick, J. M., Munger, S. C., Simecek, P., Huttlin, E. L., Choi, K., Gatti, D. M., Raghupathy, N., Svenson, K. L., Churchill, G. A., & Gygi, S. P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, *534*(7608), 500–505.

Chu, A., Robertson, G., Brooks, D., Mungall, A. J., Birol, I., Coope, R., Ma, Y., Jones, S., & Marra, M. A. (2016). Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Research*, *44*(1), e3.

Clark, D. J., Dhanasekaran, S. M., Petralia, F., Pan, J., Song, X., Hu, Y., da Veiga Leprevost, F., Reva, B., Lih, T.-S. M., Chang, H.-Y., Ma, W., Huang, C., Ricketts, C. J., Chen, L., Krek, A., Li, Y., Rykunov, D., Li, Q. K., Chen, L. S., … Clinical Proteomic Tumor Analysis Consortium. (2019). Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell*, *179*(4), 964–983.e31.

Cliff, J. M., Lee, J.-S., Constantinou, N., Cho, J.-E., Clark, T. G., Ronacher, K., King, E. C., Lukey, P. T., Duncan, K., Van Helden, P. D., Walzl, G., & Dockrell, H. M. (2013). Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *The Journal of Infectious Diseases*, *207*(1), 18–29.

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D. L., Weerasinghe, A., Huang, K.-L., Tokheim, C., Cortés-Ciriano, I., Jayasinghe, R., Chen, F., Yu, L., Sun, S., Olsen, C., Kim, J., Taylor, A. M., Cherniack, A. D., … Cancer Genome Atlas Research Network. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, *173*(2), 305–320.e10.

Edwards, B. J., Haynes, C., Levenstien, M. A., Finch, S. J., & Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. *BMC Genetics, 6*, 18.

Gale, D., & Shapley, L. S. (1962). College admissions and the stability of marriage. The American Mathematical Monthly, 69(1), 9-15.

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., 3rd, Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H., Hu, K., Andreev-Drakhlin, A. Y., Kim, J., Hess, J. M., Haas, B. J., Aguet, F., Weir, B. A., Rothberg, M. V., Paolella, B. R., … Sellers, W. R. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, *569*(7757), 503–508.

Gillette, M. A., Satpathy, S., Cao, S., Dhanasekaran, S. M., Vasaikar, S. V., Krug, K., Petralia, F., Li, Y., Liang, W.-W., Reva, B., Krek, A., Ji, J., Song, X., Liu, W., Hong, R., Yao, L., Blumenberg, L., Savage, S. R., Wendl, M. C., … Clinical Proteomic Tumor Analysis Consortium. (2020). Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell*, *182*(1), 200–225.e35.

González-Reymúndez, A., de Los Campos, G., Gutiérrez, L., Lunt, S. Y., & Vazquez, A. I. (2017). Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. *European Journal of Human Genetics: EJHG*, *25*(5), 538–544.

Jagga, Z., & Gupta, D. (2014). Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC Proceedings*, *8*(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer), S2.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* , *8*(1), 118–127.

Knights, D., Kuczynski, J., Koren, O., Ley, R. E., Field, D., Knight, R., DeSantis, T. Z., & Kelley, S. T. (2011). Supervised classification of microbiota mitigates mislabeling errors. *The ISME Journal*, *5*(4), 570–573.

Lynch, A. G., Chin, S.-F., Dunning, M. J., Caldas, C., Tavaré, S., & Curtis, C. (2012). Calling sample mix-ups in cancer population studies. *PloS One*, *7*(8), e41815.

Malossini, A., Blanzieri, E., & Ng, R. T. (2006). Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics* , *22*(17), 2114–2121.

Martín-Merino, M. (2013). A kernel SVM algorithm to detect mislabeled microarrays in human cancer samples. *13th IEEE International Conference on BioInformatics and BioEngineering*, 1–4.

Ma, X.-J., Salunga, R., Tuggle, J. T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B. M., Zhou, Y.-X., Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T. M., Erlander, M. G., & Sgroi, D. C. (2003). Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(10), 5974–5979.

Mertins, P., Mani, D. R., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J. T., Gatza, M. L., Wilkerson, M., Perou, C. M., Yellapantula, V., … NCI CPTAC. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, *534*(7605), 55–62.

Moloney, C., Rayaprolu, S., Howard, J., Fromholt, S., Brown, H., Collins, M., Cabrera, M., Duffy, C., Siemienski, Z., Miller, D., Swanson, M. S., Notterpek, L., Borchelt, D. R., & Lewis, J. (2016). RETRACTED ARTICLE: Transgenic mice overexpressing the ALS-linked protein Matrin 3 develop a profound muscle phenotype. *Acta Neuropathologica Communications*, *4*(1), 1–12.

Muhlenbach, F., Lallich, S., & Zighed, D. A. (2004). Identifying and Handling Mislabeled Instances. *Journal of Intelligent Information Systems*, *22*(1), 89–109.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics: MCP*, *1*(5), 376–386.

Prasad, V. V., & Gopalan, R. O. (2015). Continued use of MDA-MB-435, a melanoma cell line, as a model for human breast cancer, even in year, 2014. *NPJ Breast Cancer*, *1*, 15002.

Rae, J. M., Creighton, C. J., Meck, J. M., Haddad, B. R., & Johnson, M. D. (2007). MDA-MB-435 cells are derived from M14 melanoma cells--a loss for breast cancer, but a boon for melanoma research. *Breast Cancer Research and Treatment*, *104*(1), 13–19.

Rai, A. J., Zhang, Z., Rosenzweig, J., Shih, I.-M., Pham, T., Fung, E. T., Sokoll, L. J., & Chan, D. W. (2002). Proteomic approaches to tumor marker discovery: identification of biomarkers for ovarian cancer. *Archives of Pathology & Laboratory Medicine*, *126*(12), 1518–1526.

Ramalingam, P., Palanichamy, J. K., Singh, A., Das, P., Bhagat, M., Kassab, M. A., Sinha, S., & Chattopadhyay, P. (2014). Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. *RNA* , *20*(1), 76–87.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Research*, *14*(10A), 1902–1910.

Samuels, D. C., Burn, D. J., & Chinnery, P. F. (2009). Detecting new neurodegenerative disease genes: does phenotype accuracy limit the horizon? *Trends in Genetics: TIG*, *25*(11), 486–488.

Sánchez, J. S., Barandela, R., Marqués, A. I., Alejo, R., & Badenas, J. (2003). Analysis of new techniques to obtain quality training sets. *Pattern Recognition Letters*, *24*(7), 1015–1022.

Sen, P. K., & Others. (2005). Gini diversity index, Hamming distance, and curse of dimensionality. *Metron-International Journal of Statistics*, *63*(3), 329–349.

Snoek, L. B., Van der Velde, K. J., Arends, D., Li, Y., Beyer, A., Elvin, M., Fisher, J., Hajnal, A., Hengartner, M. O., Poulin, G. B., Rodriguez, M., Schmid, T., Schrimpf, S., Xue, F., Jansen, R. C., Kammenga, J. E., & Swertz, M. A. (2012). WormQTL—public archive and analysis web portal for natural variation data in Caenorhabditis spp. *Nucleic Acids Research*, *41*(D1), D738–D743.

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* , *28*(1), 112–118.

Sun, Z., Wigle, D. A., & Yang, P. (2008). Non-Overlapping and Non–Cell-Type–Specific Gene Expression Signatures Predict Lung Cancer Survival. In *Journal of Clinical Oncology* (Vol. 26, Issue 6, pp. 877–883). https://doi.org/10.1200/jco.2007.13.1516

Toker, L., Feng, M., & Pavlidis, P. (2016). Whose sample is it anyway? Widespread misannotation of samples in transcriptomics studies. *F1000Research*, *5*, 2103.

Tonon, G., Wong, K.-K., Maulik, G., Brennan, C., Feng, B., Zhang, Y., Khatry, D. B., Protopopov, A., You, M. J., Aguirre, A. J., Martin, E. S., Yang, Z., Ji, H., Chin, L., & Depinho, R. A. (2005). High-resolution genomic profiles of human lung cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9625–9630.

Valenstein, P. N., Raab, S. S., & 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. *Archives of Pathology &*

Laboratory Medicine, *130*(8), 1106–1113.

van der Velde, K. J., de Haan, M., Zych, K., Arends, D., Snoek, L. B., Kammenga, J. E., Jansen, R. C., Swertz, M. A., & Li, Y. (2013). WormQTLHD—a web database for linking human disease to natural variation data in C. elegans. *Nucleic Acids Research*, *42*(D1), D794–D801.

Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O. A., Gritsenko, M. A., Zimmerman, L. J., McDermott, J. E., Clauss, T. R., Moore, R. J., Zhao, R., Monroe, M. E., Wang, Y.-T., Chambers, M. C., … Clinical Proteomic Tumor Analysis Consortium. (2019). Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell*, *177*(4), 1035–1049.e19.

Venkataraman, S., Metaxas, D., Fradkin, D., Kulikowski, C., & Muchnik, I. (2004). Distinguishing mislabeled data from correctly labeled data in classifier design. *16th IEEE International Conference on Tools with Artificial Intelligence*, 668–672.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Mills Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120.

Wen, B., Mei, Z., Zeng, C., & Liu, S. (2017). metaX: a flexible and comprehensive software for processing metabolomics data. *BMC Bioinformatics*, *18*(1), 183.

Westra, H.-J., Jansen, R. C., Fehrmann, R. S. N., te Meerman, G. J., van Heel, D., Wijmenga, C., & Franke, L. (2011). MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* , *27*(15), 2104–2111.

Yoo, S., Shi, Z., Wen, B., Kho, S., Pan, R., Feng, H., Chen, H., Carlsson, A., Edén, P., Ma, W., Raymer, M., Maier, E. J., Tezak, Z., Johanson, E., Hinton, D., Rodriguez, H., Zhu, J., Boja, E., Wang, P., & Zhang, B. (2021). A community effort to identify and correct mislabeled samples in proteogenomic studies. *Patterns (New York, N.Y.)*, *2*(5), 100245.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J. C., Carr, S. A., … NCI CPTAC. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382–387.

Zych, K., Snoek, B. L., Elvin, M., Rodriguez, M., Van der Velde, K. J., Arends, D., Westra, H.-J., Swertz, M. A., Poulin, G., Kammenga, J. E., Breitling, R., Jansen, R. C., & Li, Y. (2017). reGenotyper: Detecting mislabeled samples in genetic data. *PloS One*, *12*(2), e0171324.

APPENDIX A

Table A: All 48 datasets collected from Expression Atlas, along with the number of mislabeled data and mislabeled assay in each dataset. A subset of the table is used to create Table 6.4, which contains only datasets with mislabeled data.

Dataset ID	Number of Subjects	Subjects with mislabeled data	Number of Assay	Mislabeled Assay
E-GEOD-100833	289	0	1653	0
E-MTAB-2232	377	0	1399	0
E-MTAB-6559	125	0	369	0
E-MTAB-6558	107	3	288	3
E-GEOD-19519	112	1	224	1
E-GEOD-20181	57	0	171	0
E-MTAB-7032	61	3	158	3
E-GEOD-41168	42	3	140	3
E-GEOD-31348	27	17	135	25
E-GEOD-58558	19	2	109	2
E-GEOD-63085	29	0	84	0
E-GEOD-11348	31	0	93	0
E-GEOD-53552	25	0	96	0
E-GEOD-11903	15	0	85	0
E-GEOD-41663	15	0	81	0
E-GEOD-23597	42	1	107	1
E-TABM-1138	142	6	284	6

E-MEXP-3756	20	0	40	0
E-MEXP-2069	20	0	60	0
E-GEOD-20489	11	0	54	0
E-MTAB-8549	27	0	54	0
E-MTAB-6212	15	0	45	0
E-TABM-740	18	0	36	0
E-MTAB-5262	10	0	35	0
E-GEOD-18995	16	0	32	0
E-GEOD-48445	15	0	30	0
E-GEOD-31652	13	0	26	0
E-MTAB-7456	15	0	30	0
E-MTAB-6555	10	0	30	0
E-MTAB-7087	8	0	23	0
E-GEOD-29908	9	0	18	0
E-GEOD-11227	8	0	16	0
E-MTAB-6556	10	0	180	0
E-TABM-271	8	0	32	0
E-MTAB-6473	8	0	32	0
E-MEXP-941	8	0	32	0
E-GEOD-26104	8	0	32	0
E-GEOD-80060	74	0	148	0
E-GEOD-60424	20	0	134	0
E-GEOD-21610	30	0	60	0

E-GEOD-32407	10	0	60	0
E-GEOD-60590	14	0	32	0
E-GEOD-22278	16	0	32	0
E-GEOD-16797	17	0	34	0
E-GEOD-46665	9	0	25	0
E-GEOD-11199	12	0	24	0
E-MEXP-1901	8	0	16	0
E-GEOD-11100	11	0	22	0
Total	1993	36 (1.81%)	6900	44 (0.64%)