Maximizing the Reusability of Public Gene Expression Data by Predicting Missing Metadata

Pei-Yau Lung[1], Xiaodong Pang[2], Yan Li[3], Jinfeng Zhang[1, *]

[1]Department of Statistics, Florida State University, Tallahassee, FL 32306
[2]Insilicom LLC, Tallahassee, FL 32303
[3]Department of Breast Surgery, Peking Union Medical College Hospital, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China

Abstract:

Reusability is part of the FAIR data principle, which aims to make data Findable, Accessible, Interoperable, and Reusable. One of the current efforts to increase the reusability of public genomics data has been to focus on the inclusion of quality metadata associated with the data. When necessary metadata are missing, most researchers will consider the data useless. In this study, we develop a framework to predict the missing metadata of gene expression datasets to maximize their reusability. We propose a new metric called Proportion of Cases Accurately Predicted (PCAP), which is optimized in our specifically-designed machine learning pipeline. The new approach performed better than pipelines using commonly used metrics such as F1-score in terms of maximizing the reusability of data with missing values. We also found that different variables might need to be predicted using different machine learning methods and/or different data processing protocols. Using differential gene expression analysis as an example, we show that when missing variables are accurately predicted, the corresponding gene expression data can be reliably used in downstream analyses.

Keywords: Gene expression data, reusability, proportion of cases accurately predicted, PCAP, predicting metadata, missing metadata, machine learning.

I. Introduction

Currently, large volumes of high-throughput genomic data are being generated in biomedical research every day by laboratories in both academia and industry. For example, as of May 23, 2018, the gene expression omnibus (GEO) database [1] consists of a total of 2,498,466 samples in 98,354 series generated by 18,519 different experimental platforms. Many federally-funded projects and initiatives are also generating unprecedented large volumes of genomic data, such as The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/), the Genotype-Tissue Expression (GTEx) project [2], and the ENCyclopedia Of DNA Elements (ENCODE) project [3]. The availability of so-called biomedical Big Data has allowed new scientific discoveries to be made by mining and analyzing such data. As high-throughput datasets are rich in information on cellular events, the same dataset can be reanalyzed alone or together with other data to address important questions that were previously not studied or not feasible due to limited availability of data [4-6].

However, the majority of public genomic data do not contain enough metadata, which severely limits their reusability. Overcoming this limited reusability forms part of the FAIR (Findable, Accessible, Interoperable and Reusable) data principle [7]. For example, to understand the heterogeneity of breast cancer and to develop personalized treatment for breast cancer [4, 8], the biomarker information that determines the subtypes of breast

cancer samples, such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status, are normally required. Information on race is also necessary to study the racial disparity of breast cancer [5, 6, 9]. We did an analysis of available breast cancer gene expression data generated by platform GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) in the GEO database and found that there are 29,631 samples, of which ~85% do not have ER information, ~88% do not have HER2 information, and ~95% do not have race information. This high percentage of missing metadata severely limits what can be studied using these data. Accurate prediction of the missing metadata will substantially increase the reusability of the existing genomic data.

There are two major strategies to predict missing metadata for existing genomic data. First, the traditional approach is to fill missing data by imputation, which uses the information of the same variable from other subjects in the same dataset to infer the missing data using relatively simple statistics [10]. For example, to fill the missing value for variable *i* for subject *j*, one can use the mean value of variable *i* from all the other subjects (or subjects similar to subjects *j* based on other variables). Second, modern approaches based on machine learning technology are now widely used for imputing missing data in biomedical sciences with better performance than the traditional approach [10-14].

It has been shown that gene expression is very informative and highly predictive of various clinical outcomes, such as the status of biomarkers [15-21], tumor types/status [22-25], the risks of recurrence [26, 27] and survival [27-30], and therapeutic response [4, 31-34]. For gene expression data, the gene expression profiles themselves, therefore, are ideally suited for inferring the missing metadata. It may seem quite straightforward to infer missing metadata using gene expression profiles, given the abundance of previous research works. However, if our goal is to recover the missing metadata and use such information to perform additional analysis, several issues need to be considered. First, we will likely want to exclude the data for which our prediction may not be accurate enough since the error will be transferred to downstream analysis. One may prefer the accuracy of the prediction to be at least above a certain threshold. Second, as genomic data are being generated by many different platforms, transferring models generated from one platform to other platforms is of concern. Third, the existing metrics for evaluating machine learning methods, such as area under the curve (AUC), F1-score, precision, recall, and accuracy, are not ideal for such tasks as they aim to optimize the accuracy for whole datasets. A better objective would be to recover as much data as possible with accuracy above the threshold one prefers. Here accuracy can be defined by any proper metrics, such as AUC, F1-score, etc.. We believe this objective is better suited to maximize the reusability of public gene expression data when predicting missing metadata.

In this study, we investigate the above issues to infer missing metadata using multiple gene expression datasets with a wide variety of machine learning methods. We evaluate their performance on a few representative clinical variables to assess the accuracy level the current machine learning methods can achieve.

We also investigate a robust normalization approach, rank normalization, on prediction performance. Models built using data generated by rank normalization are likely more transferable than other commonly used normalization methods, such as global normalization and quantile normalization [35, 36]. We introduce a new performance measure to evaluate the effectiveness of methods in recovering missing metadata, called Proportion of Cases Accurately Predicted (PCAP). PCAP is the percentage of data that can be recovered given a desired level of accuracy, where the actual accuracy measure (i.e. overall accuracy, precision, F1-score, etc.) can be defined by the researcher. $PCAP_{90}$ and $PCAP_{95}$ stand for the proportion of data that can be recovered with an accuracy of 90% and 95%, respectively. Through this study, we propose a framework to select the optimal pipeline, which includes several components such as data processing, oversampling method, variable selection, machine learning model and choice of performance measures, for recovering missing metadata by maximizing $PCAP_{90}$ or $PCAP_{95}$.

II. Materials and Methods

Data

Gene expression data from both sequencing and microarray platforms were used in this study. The RNA-seq data were obtained from TCGA. Each sample contains 20,483 RNA gene expression values and metadata, such as race, receptor status, and tumor type. We used data from five cancer types: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), and prostate adenocarcinoma (PRAD). The total number of patient samples from TCGA is 2,954. The microarray data were collected from 6 data series in the GEO database [1] generated by the GPL570 platform, which contains 879 breast cancer patient samples in total with 22,283 gene expression values in each sample.

We chose race, ER, PR, and HER2 status in the sequencing data as the four example variables from all the variables in the metadata. For each variable, the subset of samples with missing values was not considered in the analysis of the corresponding variable, since we cannot evaluate the performance of our method if we use these samples. For race, we included only African American (AA) and Caucasian American (CA) samples to simplify the analysis and discussions. For ER, PR, and HER2 variables, only samples with positive or negative values were included, while samples with “unequivocal”, “unknown”, etc. values were excluded. In TCGA data, ER, PR and HER2 status were only available for breast cancer patients. In the microarray data, in addition to ER, PR, and HER2 variables, we chose treatment response instead of race because of the unavailability of race data. There are two possible values for treatment response: pathological complete response (pCR) and residual disease (RD). Figure 1 shows the percentages of each possible value for the clinical variables used in this study. We assigned binary values to each clinical variable where we labeled the value with less number of

cases as +1 and the value with more number of cases as -1 (or 0 depending on the convention used in a machine learning algorithm). For example, for race, we labelled AA as +1 and CA as -1.

Normalization for gene expression

Three normalization methods were used in this study: read per million (RPM) normalization for next generation sequencing (NGS) data, quantile normalization for microarray data and rank normalization (RN) for both NGS and microarray data. For RN, each gene expression value was replaced by its rank, ranging from 1 (lowest) to *n* (highest), within a patient. If there was a tie for two or more values, the average of these values was used as the rank. The ranks were rescaled by dividing by the sum of total ranks and multiplying by 10^6.

Oversampling methods

There are many oversampling methods based on different assumptions and goals [37-39]. We chose the Synthetic Minority Oversampling Technique (SMOTE) [37] in our study. Rather than oversampling with replacement, SMOTE increases the number of the minority class by creating synthetic samples. These synthetic samples are generated from the information of minority samples. With the minority and synthetic samples, classifiers will build larger and more general decision regions, which tend to give more accurate predictions. SMOTE can be applied from the Imbalanced-learn package in Python [40].

Models

We define a model as a combination of a machine learning algorithm, number of genes selected, oversampling choice, and data normalization choice. The machine learning algorithms used in this study, include XGBoost [41], random forest (RF) [42], support vector machines (SVM) [43], and LASSO [44]. Tuning parameters were optimized from 10 sets of random combinations using 3-fold cross validation. To select the most predictive genes, a Welch two-sample t-test was first conducted to find differentially-expressed genes between two classes at the significance level of 0.1, where the two classes correspond to the two binary values of the variables to be predicted. Next we use recursive feature elimination (RFE) [45] to select an optimal set of genes (smaller than 100) or a predefined number (10 or 25) of genes using 3-fold cross validation. We restricted the optimal number of genes to less than or equal to 100.

Evaluation measures and optimization procedure

To evaluate the prediction performance of each model, three measures were used: Area under the receiver operating characteristic curve (AUROC), F_1-score, and our proposed new measure PCAP. In this study, $PCAP_x$ stands for the percentage of cases which can be predicted with x% precision in a model. Other performance measures other than precision can also be used in practice, such as accuracy, recall, or F1-score. The $PCAP_x$ is calculated from a 10-fold cross validation each with 90% of the data as training and 10% as testing. For each of

the 10 folds, a *k*-fold (*k*=10 in this study) cross validation was further performed resulting in a smaller training dataset and a validation dataset in each fold. The predicted probabilities of the validation set were first sorted. Each percentile of the predicted probability (from 50 to 99) was used as a cutoff to calculate the precision and recall in the validation set. If a percentile results in a precision greater than or equal to x%, than it would be labeled as 1 (otherwise 0). The smallest percentile with most 1s (up to *k*) among the *k* folds was selected as the cutoff and the corresponding average recall was recorded. After the 10-fold cross-validation, the average of the selected percentile from each *k*-fold cross-validation was chosen as the cutoff value for the test dataset, and the average recall is used as the estimated $PCAP_x$ for the test dataset.

The 10-fold cross validation in the training data helps to find the cutoff that results in the desired precision. The returned value of recall is the $PCAP_x$, which is an estimate of the recall value in the unseen test data given x% of precision. In this study, we use $PCAP_{90}$ for the assessment of different models.

Another benefit of performing cross validation in the training data is for model selection. Given various models, we select the model that has the best median PCAP, instead of the maximum PCAP, to achieve a more robust estimate of the average PCAP value. This selected model is most likely to generalize well when recovering missing metadata in future datasets.

III. Results

Predictive performance in sequencing data

The performance of models was tested by stratified 10-fold cross validation. Table 1 shows the median of AUROC values across 10 folds under each model. The AUROC value reached up to 0.987 when predicting race, and reached over 0.9 when predicting ER and PR status. When predicting HER2 status, the AUROC value reached 0.797. The F_1-scores in Table S1 (Supplementary file) confirm that these clinical variables can be predicted with good accuracy. Models with small, pre-defined number of genes can obtain good performance as well. Tables S2a and S2b show that using 10 or 25 genes results in satisfactory results. There is no single model that performs the best for all four variables.

Predictive performance in microarray data

The medians of AUROC values across 10 folds under each model (Table 2) indicate the models can also predict clinical variables well using gene expression data generated by microarray experiments. The AUROC values obtained are as high as 0.918 for pCR, 0.970 for ER, 0.947 for PR, and 0.938 for HER2. Table S3 (Supplementary file) shows the F1-scores of different models and settings. For predicting ER, PR, and HER2 status, it is possible to get comparative performance using fewer number of genes, which are shown in Tables S4a and S4b. Again, the results show there is no single model that performs the best for all four variables.

Effect of rank normalization

Gene expression data are usually generated by many different experimental platforms. Currently, at the gene expression omnibus (GEO) database [46], there are 15 major platforms measuring gene expression (mRNA values), each with more than 10,000 samples, together with hundreds of minor platforms with smaller number of samples. Using commonly used normalization methods, such as quantile and global normalization, models developed using data from one platform may not be generalizable to data produced by other platforms [47]. Developing different models for data generated by different platforms may not be an ideal solution for this problem. Rank normalization (RN) is a robust normalization method since only normalized rank information is kept in the normalization process. Most normalization methods do not change the rank (relative order) of the gene expression values within a sample. Here we investigate whether RN can be used as a robust normalization method for building machine learning models, which will yield better transferability of the resulting models.

To that end, we subtracted the F_1-score of either RPM (sequencing data) or Quantile normalization (microarray data) from the F_1-score of RN of each fold and took the average of the differences from each fold. As can be seen from Figure 2, RN helps to improve the performance of LASSO and SVM in sequencing data while giving comparable results in other cases. Predictions using rank-normalized gene expression values got higher F_1-scores than using RPM normalization (Figure 2 and Table S1). It is interesting to note that RN has comparable or better performance than commonly used normalization methods (RPM for sequencing data and Quantile normalization for microarray data) although it loses some valuable information during the normalization process.

$PCAP_{90}$ after model selection

Table 3 shows the predictive performance in terms of $PCAP_{90}$ as described in Section II. Numbers are the average $PCAP_{90}$ of 10-fold cross validation, where in each fold the model was selected by our model selection pipeline. All models were trained by maximizing either the F_1-score or $PCAP_{90}$, respectively. As shown in Table 3, models trained by optimizing $PCAP_{90}$ obtained higher $PCAP_{90}$ than optimizing F_1-score in both sequencing and microarray data.

Using predicted information in statistical inference

The goal of predicting missing metadata is to use the corresponding gene expression data in other analyses. In this section, we investigate the usefulness of the predicted data in downstream analyses. A very common task for gene expression data analysis is differential gene expression analysis (DGEA). We performed DGEA between Caucasian (CA) and African American (AA) breast cancer patients based on the race information provided in TCGA and the race information predicted using our method. Totally we have 2221 patient tissue samples who are either CA (1994) or AA (227), which was divided into ten folds with approximately equal

number of samples in each fold. The training data (Dataset1) has 8 folds, test data (Dataset2) has 1 fold, and the last fold was treated as a newly collected dataset (Dataset3, more details later). We first performed DGEA in Dataset1 by randomly sampling 1 fold of data (with the same CA:AA ratio) to generate the list of differentially expressed genes (DEGs), called DEG1. We then conducted DGEA for Dataset2 using their true race information to obtain another list of DEGs, called DEG2_t. The third DEGA was performed using Dataset2 again by using predicted race, where the predictive model was built using Dataset1. This gave DEG2_p. The fourth DGEA was done using Dataset3 to produce DEG3. In this comparison, we assume that we can collect some new gene expression data with race information and perform the same DGEA. The last DGEA was done by first randomly permuting the race in Dataset2 to generated DEG2_r. We compute the number of overlapping DEGs between DEG1 and DEG2_t, between DEG1 and DEG2_p, between DEG1 and DEG3, and between DEG1 and DEG2_r. Computing the number of overlapping DEGs was done iteratively for each of the ten folds, then the average number of overlapping DEGs was computed. We repeated the process 50 times and drew the box-plot for the average number of overlapping DEGs in Figure 3.

The box-plots and significance tests showed that when computing the number of overlapping DEGs, there is no significant difference between using true race information and using predicted race. There is no significant difference between using predicted race and using the true race of newly collected data, either. This indicates that datasets with metadata predicted using our method can be used reliably in other analysis.

IV. Discussion

Genomics data are not useful if necessary metadata are not available. Unfortunately, a large amount of such important metadata is missing in public genomics datasets. In this study, we proposed a framework for maximizing the reusability of public gene expression data by predicting the missing metadata using machine learning methods. To develop and validate the framework, we used microarray and sequencing gene expression data with a total of 3,833 cancer patient tissue samples to investigate whether we can predict missing metadata, such as race, ER, PR, HER2 and treatment response. Our study has shown that gene expression profiles can be used to predict metadata accurately. Out of over 20,000 genes, we can select small numbers of genes to obtain reliable predictions. For those variables for which reliable predictions cannot be achieved for all the missing metadata, we can select a subset of reliable predictions using our pipeline and a new measure designed for maximizing the reusability of public gene expression data. We found that different variables require different methods and parameter settings to achieve optimal performance. This is consistent with the well-known notion that no single method is the best for all kinds of machine learning tasks.

In addition to machine learning algorithms, normalization methods can have a substantial effect on the performance as well. In this study, we found that the robust rank normalization (RN) [48, 49] can produce better or comparable performance than commonly used normalization methods (RPM for sequencing data and

quantile normalization for microarray data). Predictions from rank-normalized sequencing data resulted in higher F_1-scores. In microarray data, RN gave comparable performances. RN is less sensitive to experimental noise and outliers. It also allows researchers to conveniently combine several datasets generated from different platforms. Our result indicates that RN can be a good choice when building machine learning models either for predicting metadata to maximize the reusability of public gene expression data or for general model building purposes using gene expression data.

Successful statistical models depend on the quality of the data used for building the models. The accuracy of the predicted values for the missing data needs to be carefully evaluated to ensure the quality of the data to be used in downstream applications. While traditional evaluation metrics can be used, they are not ideal because they aim for accurate predictions of whole datasets. When researchers try to reuse public genomics data, they do not need to use all the data. Additionally, the accuracy for the whole dataset may not reach the desired accuracy threshold. They can use only the part of the data for which missing metadata can be reliably predicted. Here, a reasonable metric to optimize would be, give a certain accuracy threshold, the proportion of data that can be predicted above that threshold. The higher the number, the more data we will have for reuse with desired accuracy.

With the above reasoning, we proposed the metric Proportion of Cases Accurately Predicted (PCAP) for the purpose of maximizing reusability of public gene expression data. In addition, we proposed a selection pipeline to select a model from various combinations of algorithms, normalization methods, and data balancing procedures. Our results showed that we were able to recover a high percentage of samples with the desired accuracy. It is also recommended that one should maximize PCAP, instead of traditional performance measures for the whole dataset, when building models to obtain higher percentages of usable samples.

We also demonstrated the effectiveness of the predicted metadata in downstream inference tasks. In the study, we performed differential gene expression analysis (DGEA) using predicted race and found that the effectiveness of the analysis using predicted metadata is similar to that using true metadata. This demonstrated that our framework for maximizing the reusability of gene expression data can be reliably used in the future by other researchers.

There is a possibility that the subset of the data that can be predicted with high accuracy is systematically different in some way from the whole dataset. In such cases, the conclusion one can draw will be limited to only the subset of the data that can be predicted with high accuracy. In real clinical settings, for example biomarker discoveries, one can use the predictive model to stratify patient population and only apply the discoveries (i.e. identified diagnostic biomarkers) to that population.

It is worth noting that the metadata we have used as true information may have some noise in them. For example, it is well acknowledged that the self-reported race/ethnicity has high inaccuracy levels [50, 51]. These inherent errors will limit the upper bound of the accuracy we can achieve.

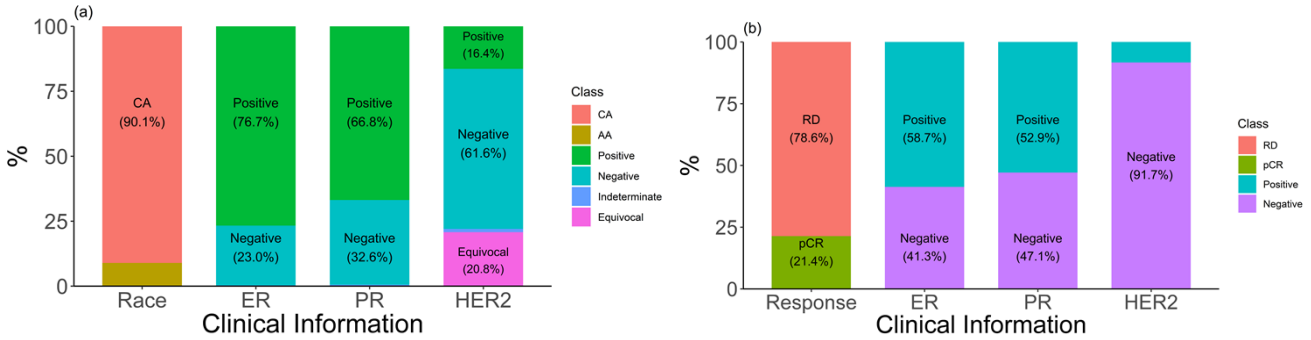

Figure 1. The number of patients in each class of four clinical variables for both sequencing data from TCGA (a) and microarray data from GEO (b). Numbers in parenthesis are the percentages of the corresponding class. AA: African American. CA: Caucasian American. pCR: pathological complete response. RD: residue disease.

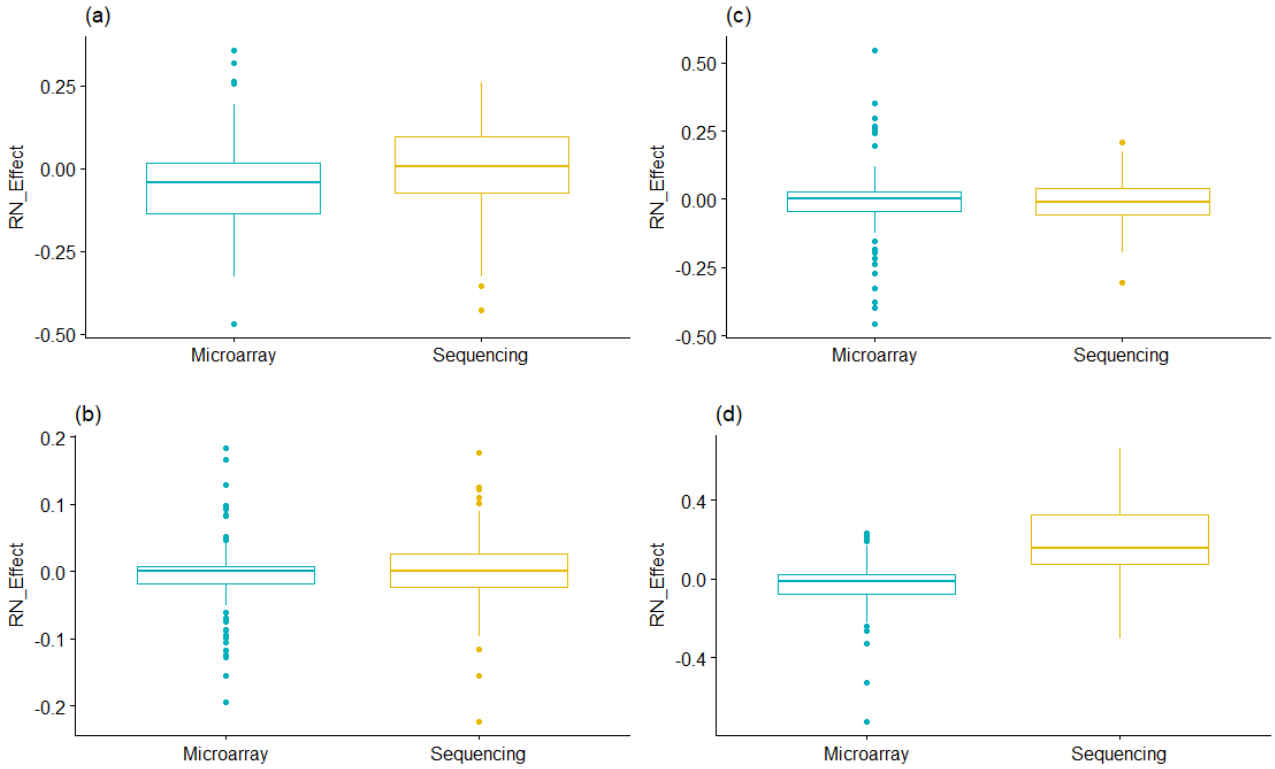

Figure 2. Comparison of rank normalization (RN). RN_Effect denotes the F_1-score from RN minus F_1-score from RPM of each fold using: (a) LASSO, (b) Random Forest, (c) XGBoost, and (d) SVM. For LASSO and SVM, RN gave better performance for sequencing data. For other models, RN gave comparable performances.

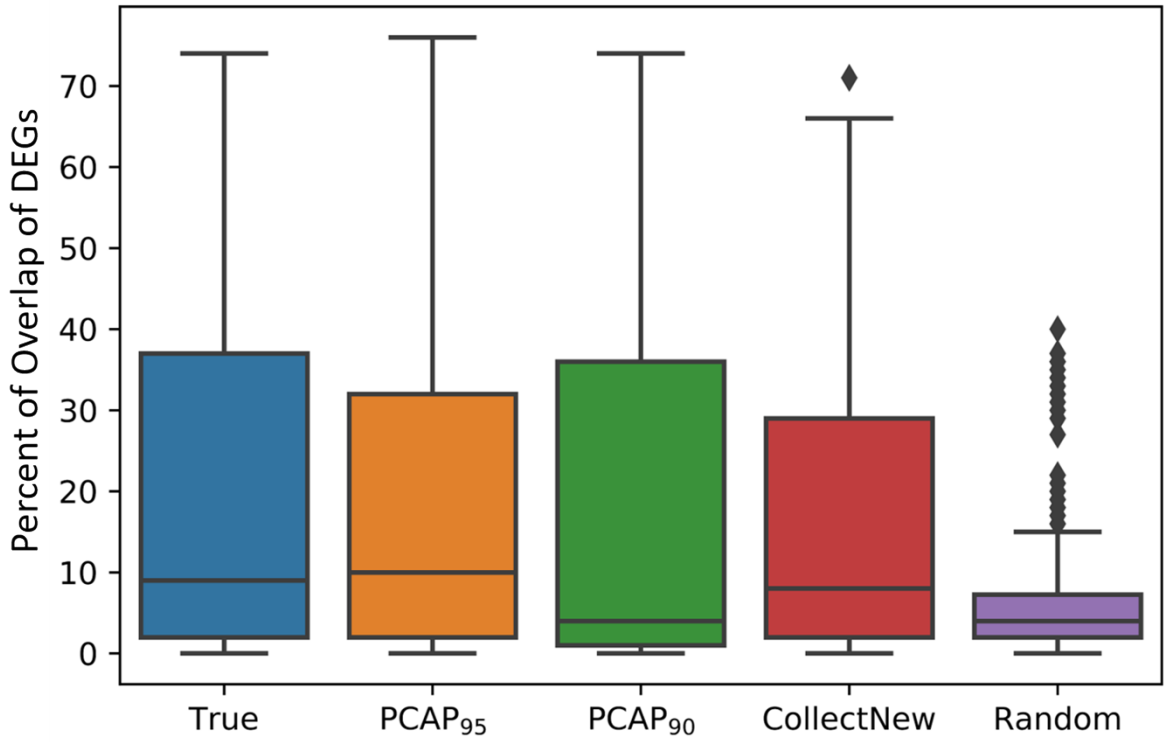

Figure 3. Box plot for the overlap of differential expressed genes (DEGs) detected by differential gene expression analyses using five different datasets. The numbers are averages of ten runs. True: DEGs obtained using the true race information in the original TCGA data; $PCAP_{95}$: DEGs obtained using a dataset where race is predicted by optimizing $PCAP_{95}$; $PCAP_{90}$: DEGs obtained using a dataset where race is predicted by optimizing $PCAP_{90}$; CollectNew: DEGs obtained by collecting a new dataset with known race information. This dataset is part of the TCGA data which was not used in other analyses; Random: DEGs by randomly permuting the race assignments among the patients in the dataset.

Table 1. AUROC values from different algorithms for variables in sequencing data. RN denotes that the predictors were rank-normalized. SMOTE denotes that the samples were balanced by Synthetic Minority OverSampling Technique. An optimal number of genes were selected by recursive feature elimination with cross validation.

	Race			**ER**[a]		
	RPM	RN	RN + SMOTE	RPM	RN	RN + SMOTE
LASSO	0.946	0.979	0.982	0.895	0.915	0.906
Random Forest	0.946	0.931	0.975	0.949	0.949	0.957
XGBoost	0.982	0.975	0.985	0.947	0.947	0.948
SVM	0.835	0.987	0.975	0.834	0.908	0.904
	PR[b]			**HER2**[c]		
LASSO	0.836	0.883	0.875	0.701	0.791	0.785
Random Forest	0.909	0.920	0.902	0.788	0.793	0.797
XGBoost	0.899	0.900	0.899	0.793	0.784	0.788
SVM	0.789	0.858	0.863	0.702	0.792	0.790

a: Only ER-positive (833) or ER-negative (243) patients were included.
b: Only PR-positive (726) or PR-negative (347) patients were included.
c: Only HER2-positive (173) or HER2-negative (581) patients were included.

Table 2. AUROC values from different algorithms for variables in microarray data.

	Response			**ER**[a]		
	RPM	RN	RN + SMOTE	RPM	RN	RN + SMOTE
LASSO	0.910	0.899	0.875	0.954	0.960	0.959
Random Forest	0.918	0.909	0.914	0.968	0.970	0.968
XGBoost	0.884	0.883	0.899	0.968	0.967	0.960
SVM	0.904	0.889	0.892	0.952	0.955	0.960
	PR[b]			**HER2**[c]		
LASSO	0.938	0.939	0.941	0.938	0.924	0.901
Random Forest	0.937	0.936	0.942	0.916	0.906	0.894
XGBoost	0.938	0.938	0.933	0.894	0.880	0.917
SVM	0.947	0.928	0.927	0.931	0.909	0.896

a: Only ER-positive (833) or ER-negative (243) patients were included.
b: Only PR-positive (726) or PR-negative (347) patients were included.
c: Only HER2-positive (173) or HER2-negative (581) patients were included.

Table 3. $PCAP_{90}$ after model selection (mean). The numbers are the average of 10-fold cross-validation. By optimizing $PCAP_{90}$, we can generally achieve better $PCAP_{90}$ values (and/or better precision) compared to optimizing F1-scores. In all cases, by optimizing $PCAP_{90}$, we can achieve the desired precision of 90%. However, by optimizing F1-scores, this cannot be always achieved (colored as red).

Sequencing data					
	Race	ER	PR	HER2	**Average**
Optimizing F_1-score	0.217(0.863)	0.167(0.858)	0.385(0.903)	0.248(0.983)	**0.254 (.902)**
Optimizing $PCAP_{90}$	0.216(0.98)	0.210(0.90)	0.360(0.962)	0.260(0.915)	**0.262 (.939)**
Microarray data					
	Response	ER	PR	HER2	**Average**
Optimizing F_1-score	0.056(0.85)	0.686(0.926)	0.570(0.953)	0.220(0.95)	**0.383 (0.920)**
Optimizing $PCAP_{90}$	0.09(0.95)	0.695(0.954)	0.553(0.966)	0.250(0.95)	**0.397 (0.955)**

Acknowledgement

JZ is supported partially by a grant from National Institute of General Medical Science of National Institute of Health, grant # R01GM126558.

References

1. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D991-5.
2. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**(7675): p. 204-213.
3. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
4. Yu, K., et al., *Personalized chemotherapy selection for breast cancer using gene expression profiles.* Sci Rep, 2017. **7**: p. 43294.
5. Shi, Y., et al., *Integrative Comparison of mRNA Expression Patterns in Breast Cancers from Caucasian and Asian Americans with Implications for Precision Medicine.* Cancer Res, 2017. **77**(2): p. 423-433.
6. Li, Y., et al., *Tumoral expression of drug and xenobiotic metabolizing enzymes in breast cancer patients of different ethnicities with implications to personalized medicine.* Sci Rep, 2017. **7**(1): p. 4747.
7. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018.
8. Lee, C.H. and H.-J. Yoon, *Medical big data: promise and challenges.* Kidney research and clinical practice, 2017. **36**(1): p. 3.
9. Stewart, P.A., et al., *Differentially expressed transcripts and dysregulated signaling pathways and networks in African American breast cancer.* PLoS One, 2013. **8**(12): p. e82460.
10. Cismondi, F., et al., *Missing data in medical databases: Impute, delete or classify?* Artificial intelligence in medicine, 2013. **58**(1): p. 63-72.
11. Hu, Z., et al., *Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record.* Journal of biomedical informatics, 2017. **68**: p. 112-120.
12. Jerez, J.M., et al., *Missing data imputation using statistical and machine learning methods in a real breast cancer problem.* Artificial intelligence in medicine, 2010. **50**(2): p. 105-115.
13. Liao, S.G., et al., *Missing value imputation in high-dimensional phenomic data: imputable or not, and how?* BMC bioinformatics, 2014. **15**(1): p. 346.
14. Nanni, L., A. Lumini, and S. Brahnam, *A classifier ensemble approach for the missing feature problem.* Artificial intelligence in medicine, 2012. **55**(1): p. 37-50.
15. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-536.
16. West, M., et al., *Predicting the clinical status of human breast cancer by using gene expression profiles.* Proceedings of the National Academy of Sciences, 2001. **98**(20): p. 11462-11467.
17. Finak, G., et al., *Stromal gene expression predicts clinical outcome in breast cancer.* Nature medicine, 2008. **14**(5): p. 518.
18. Shipp, M.A., et al., *Diffuse large B-cell lymphoma outcome prediction by gene-expression*

profiling and supervised machine learning. Nature medicine, 2002. **8**(1): p. 68-74.
19. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* science, 1999. **286**(5439): p. 531-537.
20. Dai, X., A. Chen, and Z. Bai, *Integrative investigation on breast cancer in ER, PR and HER2-defined subgroups using mRNA and miRNA expression profiling.* Scientific reports, 2014. **4**: p. 6566.
21. Bastani, M., et al., *A machine learned classifier that uses gene expression data to accurately predict estrogen receptor status.* PloS one, 2013. **8**(12): p. e82144.
22. Wei, J.S., et al., *Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma.* Cancer Res, 2004. **64**(19): p. 6883-91.
23. Pomeroy, S.L., et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**(6870): p. 436.
24. Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature medicine, 2001. **7**(6): p. 673.
25. Dudoit, S., J. Fridlyand, and T.P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data.* Journal of the American statistical association, 2002. **97**(457): p. 77-87.
26. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer cell, 2002. **1**(2): p. 203-209.
27. Smith, J.J., et al., *Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer.* Gastroenterology, 2010. **138**(3): p. 958-968.
28. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature medicine, 2002. **8**(8): p. 816.
29. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* The Lancet, 2005. **365**(9460): p. 671-679.
30. Shedden, K., et al., *Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study.* Nature medicine, 2008. **14**(8): p. 822.
31. Chang, J.C., et al., *Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer.* The Lancet, 2003. **362**(9381): p. 362-369.
32. Yu, K., et al., *Personalized chemotherapy selection for breast cancer using gene expression profiles.* Scientific reports, 2017. **7**: p. 43294.
33. Ghadimi, B.M., et al., *Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy.* Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 2005. **23**(9): p. 1826.
34. Iwao-Koizumi, K., et al., *Prediction of docetaxel response in human breast cancer by gene expression profiling.* Journal of clinical oncology, 2005. **23**(3): p. 422-431.
35. Liu, Y., J. Zhang, and X. Qiu, *Super-delta: a new differential gene expression analysis procedure with robust data normalization.* BMC Bioinformatics, 2017. **18**(1): p. 582.
36. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome biology, 2010. **11**(3): p. R25.
37. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
38. Han, H., W.-Y. Wang, and B.-H. Mao, *Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning.* 2005. **3644**: p. 878-887.
39. Nguyen, H.M., E.W. Cooper, and K. Kamei, *Borderline over-sampling for imbalanced data classification.* International Journal of Knowledge Engineering and Soft Data Paradigms, 2011. **3**(1): p. 4-21.
40. Lemaitre, G., F. Nogueira, and C.K. Aridas, *Imbalanced-learn: A python toolbox to tackle the*

curse of imbalanced datasets in machine learning. arXiv preprint arXiv:1609.06570, 2016.
41. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.
42. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
43. Steinwart, I. and A. Christmann, *Support vector machines*. 2008: Springer Science & Business Media.
44. Tibshirani, R., *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 1996: p. 267-288.
45. Guyon, I., et al., *Gene selection for cancer classification using support vector machines.* Machine learning, 2002. **46**(1-3): p. 389-422.
46. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Res, 2002. **30**(1): p. 207-10.
47. Rudy, J. and F. Valafar, *Empirical comparison of cross-platform normalization methods for gene expression data.* BMC Bioinformatics, 2011. **12**: p. 467.
48. Szabo, A., et al., *Variable selection and pattern recognition with gene expression data generated by the microarray technology.* Mathematical Biosciences, 2002. **176**(1): p. 71-98.
49. Tsodikov, A., A. Szabo, and D. Jones, *Adjustments and measures of differential expression for microarray data.* Bioinformatics, 2002. **18**(2): p. 251-260.
50. Mersha, T.B. and T. Abebe, *Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities.* Human genomics, 2015. **9**(1): p. 1.
51. Saunders, C.L., et al., *Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey.* BMJ open, 2013. **3**(6): p. e002882.