A peer-reviewed version of this preprint was published in PeerJ on 21 January 2016.

View the peer-reviewed version (peerj.com/articles/1621), which is the preferred citable publication unless you specifically need to cite this preprint.

Thompson JA, Tan J, Greene CS. 2016. Cross-platform normalization of microarray and RNA-seq data for machine learning applications. PeerJ 4:e1621 https://doi.org/10.7717/peerj.1621

Cross-platform normalization of microarray and RNA-seq data for machine learning applications

Jeffrey A Thompson, Jie Tan, Casey S Greene

Large, publicly available gene expression datasets are often analyzed with the aid of machine learning algorithms. Although RNA-seq is increasingly the technology of choice, a wealth of expression data already exist in the form of microarray data. If machine learning models built from legacy data can be applied to RNA-seq data, larger, more diverse training datasets can be created and validation can be performed on newly generated data. We developed Training Distribution Matching (TDM), which transforms RNA-seq data for use with models constructed from legacy platforms. We evaluated TDM, as well as quantile normalization and a simple $\log_2$ transformation, on both simulated and biological datasets of gene expression. Our evaluation included both supervised and unsupervised machine learning approaches. We found that TDM exhibited consistently strong performance across settings and that quantile normalization also performed well in many circumstances. We also provide a TDM package for the R programming language.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

Cross-platform normalization of microarray and RNA-seq data for machine learning applications

Jeffrey A. Thompson[1,2], **Jie Tan**[1,3], **and Casey S. Greene**[1,4,5,6]

[1]**Department of Genetics, Geisel School of Medicine at Dartmouth**
[2]**Quantitative Biomedical Sciences Program, Geisel School of Medicine at Dartmouth**
[3]**Molecular and Cellular Biology Program, Geisel School of Medicine at Dartmouth**
[4]**Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine. University of Pennsylvania**
[5]**Institute for Biomedical Informatics. Perelman School of Medicine, University of Pennsylvania**
[6]**Institute for Translational Medicine and Therapeutics, Perelman School of Medicine. University of Pennsylvania**

ABSTRACT

Large, publicly available gene expression datasets are often analyzed with the aid of machine learning algorithms. Although RNA-seq is increasingly the technology of choice, a wealth of expression data already exist in the form of microarray data. If machine learning models built from legacy data can be applied to RNA-seq data, larger, more diverse training datasets can be created and validation can be performed on newly generated data. We developed Training Distribution Matching (TDM), which transforms RNA-seq data for use with models constructed from legacy platforms. We evaluated TDM, as well as quantile normalization and a simple log_2 transformation, on both simulated and biological datasets of gene expression. Our evaluation included both supervised and unsupervised machine learning approaches. We found that TDM exhibited consistently strong performance across settings and that quantile normalization also performed well in many circumstances. We also provide a TDM package for the R programming language. A TDM R package is available at: https://github.com/greenelab/TDM (doi:10.5281/zenodo.32852).

Keywords: TDM, cross-platform, normalization, RNA-seq, microarray, machine learning, quantile normalization, training, distribution

INTRODUCTION

A wealth of gene expression data is being made publicly available by consortia such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Network et al., 2012). Such large datasets provide the opportunity to discover signals in gene expression that may not be apparent with smaller sample sizes, such as prognostic indicators or predictive factors, particularly for subsets of patients. However, discerning the signal in such large datasets frequently relies on the application of machine learning algorithms to identify relationships in high-dimensional data, or to cope with the computational complexity.

These approaches often construct a model that captures relevant features of a dataset, and the model can be used to make predictions about new data, such as how well a patient will respond to a particular treatment (Geeleher et al.), or whether their cancer is likely to recur (Kourou et al., 2014). Therefore, the model is usually constructed using a large, diverse dataset and is then applied to incoming cases to make predictions about them.

Increasingly, investigators are measuring gene expression with RNA-seq. Despite its higher cost, several advantages of RNA-seq over DNA microarrays are typically cited (Wang et al., 2010):

- RNA-seq does not require *a priori* knowledge of gene sequence.
- RNA-seq is able to detect single nucleotide variations (Atak et al., 2013).

- RNA-seq has a much higher dynamic range.
- RNA-seq provides quantitative expression levels.
- RNA-seq provides isoform-level expression measurements.

While RNA-seq represents a substantial technological advance, microarrays are still widely used because they are less expensive, are more consistent with historical data, and robust statistical methods exist for working with them. Perhaps more importantly, there are a tremendous number of historical microarray experiments that have already been performed. ArrayExpress, a publicly available database of experiments maintained by the European Bioinformatics Institute (EBI) (Rustici et al.), contains more than 60000 experiments and 1.8 million assays. As the transition to RNA-seq continues, the massive collection of microarray data constitute a rich resource of gene expression data. Therefore, training a classifier on large datasets created from microarrays and testing that classifier on samples measured with RNA-seq would be useful because new data could be generated with the most advanced technology and still be used for validation.

Machine learning models benefit from large, diverse training datasets in order to build generalizable models. However, most algorithms operate under the assumption that the training and test data will be drawn from the same distribution. When the distribution of training and test datasets differ, it can result in reduced fit of the model. This is referred to as dataset shift. Although some methods exist for machine learning under certain types of dataset shift (some of these are reviewed by (Moreno-Torres et al.)), there are no general solutions for the type of dataset shift that occurs between different gene expression platforms. In this case,

$$P_{train}(y|x) \neq P_{test}(y|x) \wedge P_{train}(x) \neq P_{test}(x) \tag{1}$$

where y is the class of the example and x is an expression value. This is in the category of "other types of dataset shift" mentioned by Moreno-Torres et al. for which there is no known general solution. It refers to the fact that the probability of the dependent variable may not be the same in the training and test set for a given value of an independent variable and that the probability of that value occurring is different in both datasets.

Normalization and batch correction techniques, such as quantile normalization, help to deal with some dataset shifts (Bolstad et al.). Although quantile normalization was developed specifically for microarrays it has also come to be widely used for RNA-seq (Wei et al., 2014; Norton et al., 2013), as well as cross-platform normalization (Li et al., 2015; Forés-Martos et al., 2015). The only method we are aware of that has been expressly designed for comparing microarrays and RNA-seq (apart from our own) is the recently published Probe Region Expression estimation Based on Sequencing (PREBS) (Uziela and Honkela, 2015). This method estimates RNA-seq expression values at microarray probe regions in order to make the data more compatible. However, the increase in comparability means discarding the expression information contained in other reads. Additionally, because this method requires access to raw reads, it cannot be used on public data where there may be privacy concerns. Thus PREBS cannot be as widely applied to publicly available data as our method or quantile normalization, which only need estimated transcript abundances and do not require transcripts and probes to match.

Given the differences in dynamic range between microarrays and RNA-seq and the fact that microarrays represent relative expression and RNA-seq quantitative counts, it may appear that the data are incommensurable. Indeed, in some cases the effect sizes for certain treatments can be dependent on the platform used (Wang et al., 2014). However, a number of papers have compared expression values from tissue samples for which both microarray and RNA-seq data have been collected. In each case, it was found that microarray and RNA-seq data are well correlated (Wang et al., 2014; Mooney et al.; Malone and Oliver), although this correlation was stronger for the more highly expressed genes. Therefore, the potential for machine learning being applied cross platform should exist, given sufficient similarity in the data distributions.

Aside from other normalization techniques that might be used, microarray data are generally analyzed after log transformation, so that values represent fold-change and statistical tests requiring normality can be used. Therefore, one possible approach to integrating microarray and RNA-seq data in a machine

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

learning pipeline would be to simply log transform the RNA-seq data. In this work, we demonstrate that this approach is insufficient to achieve consistent predictions.

In this paper, we describe Training Distribution Matching (TDM), an approach that normalizes RNA-seq data to allow models trained on microarray data to be tested on RNA-seq. We consider this approach in conjunction with two existing approaches: quantile normalization, and simple log transformation of the RNA-seq data. We compare performance on both simulated data and two different gene expression datasets from TCGA that contain both microarray and RNA-seq expression data. Finally, we examine how these methods perform using a model trained on a distinct microarray breast cancer dataset.

The intuition behind TDM is to transform the RNA-seq data so that its distribution is closer to the training data but to leave between-sample relationships intact. It aims to correct the dataset shift between the microarray and RNA-seq data, using a light touch.

We evaluated all three approaches using both unsupervised and supervised machine learning methods. For an unsupervised approach we used PAM (Kaufman and Rousseeuw, 1990) and for a supervised approach we used LASSO logistic regression (Tibshirani, 1996).

Interestingly, both TDM and quantile normalization perform well, suggesting that legacy datasets may be quite useful for such analyses. However, TDM tends to hold up better than quantile normalization in cases of increased noise in the data.

1 METHODS

The basis of our approach is to adjust the distribution of RNA-seq data to improve recognition for features learned from microarrays. Most machine learning algorithms that are applicable to expression data assume the test data are drawn from the same probability distribution as the training data. If a normalization approach makes the distributions similar but does not preserve internal data dependencies, then the model will fit poorly.

Our Training Distribution Matching (TDM) approach transforms test data to have approximately the same distribution of expression values as the training data, without changing the rank order of most genes in terms of expression levels. In other words, our method is not intended to improve the rank correlation of the datasets, since this can mean changing the biological significance of the data (particularly for RNA-seq data) and brings the validity of the results into question. Instead, it is intended to improve the recognizability of features. The distribution is adjusted for the test dataset as a whole, rather than by individual sample, to avoid over-normalization.

It is to be expected that many genes will have a different rank order between datasets, regardless of the platforms used. However, by making the expression values generally more similar between datasets, the ability of a model to fit the data will be improved. Because microarray data are generally worked with as log_2 transformed values, either the RNA-seq data must be log_2 transformed as well, or the microarray data must not. In this work we have chosen to log_2 transform the RNA-seq data, because microarray data are usually received in this form, but the package allows either decision to be made.

1.1 The Training Distribution Matching (TDM) Algorithm for Cross-platform Normalization

TDM is a normalization method that aims to make RNA-seq data comparable with microarray data without having a large effect on inter-observation dependencies. It is performed as described in Algorithm 1.

TDM establishes a relationship between the spread of the middle half of the the training data and the extremal values, then transforms the test data to have that same relationship. It determines the ratio of the spread above the third quartile to the IQR of the training data and then uses this to bound the maximum value in the testing data (i.e. it determines the number of IQRs that can be fit between the third quartile and the maximum value). The equivalent is done for the ratio of the spread below the first quartile and the IQR of the training data, but this value is not allowed below zero. Finally, each value is mapped into a range from the minimum of the training data to the maximum of the training data (in the inverse-log space) and log_2 transformed.

1.2 Quantile Normalization

Quantile normalization makes it possible to ensure that two datasets are drawn from the same distribution. Given a reference distribution, a target distribution is normalized by replacing each of its values by the

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

Algorithm 1 TDM Algorithm – *q3(S)* yields the third quartile of a set *S*, *q1(S)* yields the first quartile of *S*, *iqr(S)* yields the inter-quartile range of *S*, *max(S)* yields the maximum value in *S*, and *min(S)* yields the minimum value in *S*. *Testing* and *Training* are sets containing all expression values for all respective samples where each member is the expression value of a single gene for a single sample.

$\Delta \leftarrow \frac{2^{max(Training)} - 2^{q3(Training)}}{2^{iqr(Training)}}$
$\Delta' \leftarrow \frac{2^{q1(Training)} - 2^{min(Training)}}{2^{iqr(Training)}}$
$\sqcup \leftarrow q3(Testing) + \Delta \times iqr(Testing)$
$\sqcap \leftarrow q1(Testing) - \Delta' \times iqr(Testing)$
for $x \in Testing$ **do**
  if $x > \sqcup$ **then**
    $x \leftarrow \sqcup$
  else if $x < \sqcap$ **then**
    $x \leftarrow \sqcap$
  end if
  if $x < 0$ **then**
    $x \leftarrow 0$
  end if
  $x \leftarrow \frac{x - \sqcap}{\sqcup - \sqcap} \times (2^{max(Training)} - 2^{min(Training)}) + 2^{min(Training)}$
end for

value of the variable with the same rank in the reference distribution. If the reference distribution contains multiple samples, the target and reference distributions will only be identical if the reference distribution is first quantile normalized across all samples.

All quantile normalization was performed using the `nomalize.quantiles.use.target` method of the `preprocessCore` package (Bolstad, 2015) in the R statistical environment (R Core Team, 2015).

1.3 Evaluation

We evaluated the performance of our methods using both an unsupervised and a supervised machine learning algorithm. The unsupervised approach we chose was Partitioning Around Mediods (PAM). For PAM, we constructed a simulated dataset, so that we could observe the effect of TDM under controlled conditions. The supervised approach chosen for evaluation was LASSO multinomial logistic regression.

1.3.1 Partitioning Around Mediods (PAM)

PAM is a clustering algorithm that identifies "mediods" or examples in the dataset that represent the best centers for a user-defined number of clusters. The model that is built can be applied to new data to determine which mediod (and thus which cluster) the new data best fit to. It is similar to the k-means algorithm but tends to be more robust to outliers. Here we used the `pam` method from the `cluster` package (Maechler et al., 2015) in R.

1.3.2 LASSO multinomial logistic regression.

The supervised method we chose is LASSO multinomial logistic regression. For this method we relied on the `glmnet` package in the R (Friedman et al., 2010). A detailed description of the method can be found in (Tibshirani, 1996). We evaluated the performance on classification of either tumor subtype or class depending on the dataset. In each case, we used 100 fold cross-validation to train the model. We then assessed performance of normalization methods by applying the model constructed on one platform to a dataset from a distinct platform normalized with different approaches. This process was repeated 10 times, with different random seeds.

LASSO logistic regression builds a particularly efficient model of features, using only the variables that are the most informative (Liang et al.). It is a popular technique for selecting a sparse set of predictors in biological datasets (e.g. identifying the smallest set of genes that reliably predict if someone would benefit from a particular therapy). It can also be used for multinomial logistic regression, for cases in which multiple classifications are being considered. LASSO optimization is similar to normal regression, but it tends to reduce many of the coefficients of predictors to 0, leaving a relatively small set of predictors

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

that are best able to predict an example's class, removing redundant predictors, and leading to a model that is easier to interpret than some other approaches. This makes it particularly useful for problems like the construction of biomarkers.

1.4 Data

1.4.1 Simulated Data

The simulated data were generated using the program SynTReN (Van den Bulcke et al., 2006). This tool enables the generation of datasets that have a distribution similar to typical microarray data, with classes in the data that are differentially expressed due to some condition, and that contain correlations between gene pairs that more realistically simulate the complexity of biological data. We generated a dataset using the default settings with the following exceptions: we generated 500 genes, half of which would be background genes mostly unaffected by changing conditions; we asked for 400 samples; and we set 4 experimental conditions to be encoded in the data. Each condition received 100 samples. An additional 400 samples were created by duplicating these samples by taking the inverse log of them, rounding the results, and rescaling them to the range [0,1000000] to simulate the higher dynamic range of RNA-seq data. Although an imperfect simulation, most of what we wanted to capture is the effect of noise on datasets with matching samples but different dynamic ranges.

Additional noisy datasets were created using the `addNoise` method from the `sdcMicro` package (Templ et al., 2015) in R on the simulated datasets by adding a percentage of gaussian noise from 0 to 9.5% in increments of .5%. Each simulated RNA-seq dataset was normalized using TDM (with the simulated microarray data as a reference), quantile normalization (with the simulated microarray data as a target), or log transformation.

1.4.2 Biological Data

We used three biological datasets:

- Dataset 1 – The first contains gene expression values for tumor and tumor-adjacent normal biopsies of breast cancer from TCGA (The Cancer Genome Atlas) measured by both microarray (Agilent 244K platform) and RNA-seq (Illumina HiSeq platform). The microarray dataset contains 531 cancer samples and 63 tumor-adjacent normal samples. However, only 516 of the cancer samples and 58 of the tumor-adjacent normal samples had complete subtype data for this work, so only those were retained. The RNA-seq data included 1095 cancer samples and 113 normal. However, only 844 tumor samples and 107 normal had complete subtype data. These samples overlap 509 cancer and 60 normal samples from the microarray data. Therefore, they can be thought of as a low noise dataset for comparing results between microarray and RNA-seq. For these data, breast cancer subtype was used for classification (Cancer Genome Atlas Network et al., 2012).

- Dataset 2 – The second biological dataset contains gene expression values for tumor and tumor-adjacent normal biopsies of colon and rectal cancer from TCGA measured on the same two platforms. The microarray dataset contains 220 cancer samples and 22 normal samples, all of which were retained. The RNA-seq data included 380 cancer samples and 50 tumor-adjacent normal samples. Of these, 330 cancer samples and 29 normal included complete tumor class data and did not overlap the microarray data. In this instance, the lack of overlap was used to create a higher noise dataset. For these data Cpg island methylator phenotype (CIMP) status (Sánchez-Vega et al., 2015) was used for classification.

- Dataset 3 – The third biological dataset is based on a breast cancer compendium created in previous work (Tan et al., 2015). It again contains both microarray and RNA-seq data. However, the first microarray dataset is from METABRIC, a retrospective cohort built from tumor banks in the UK and Canada (Curtis et al., 2012) using the Illumina HT-12 v3 platform. Missing values were imputed in these data using KNNImputer from the Sleipner library (Huttenhower et al.) using 10 neighbors as recommended by (Troyanskaya et al.). These were filtered by median absolute deviation (MAD), keeping the 3000 genes with the highest MAD values. Of these genes, only 2520 were included in the TCGA microarray breast cancer data mentioned above, and so the METABRIC data were further filtered to include those 2520 genes. The RNA-seq data were also from the first breast cancer dataset but were filtered to include only the same 2520 genes and to include only the overlapping samples with the microarray data. This dataset allows us to compare microarray and

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

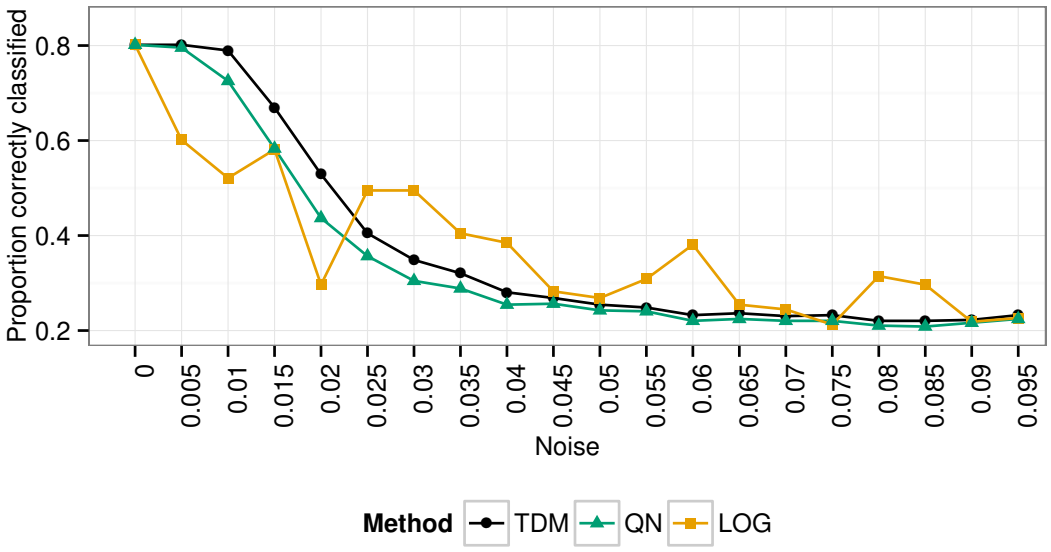


Figure 1. The proportion of samples correctly classified in the simulated data at increasing levels of noise. This is taken to be the proportion of samples clustered in a group for which the most common class matches their own. The x-axis represents increasing noise in the data. As the noise increases, the TDM transformed data consistently have the better performance than quantile normalized data. Log transformation results in erratic performance with these data.

RNA-seq data across research consortia on a set of genes selected for high variance. Furthermore, it allows us to compare the performance of normalized RNA-seq data to microarray data for the same samples.

RNA-seq and clinical data were obtained from the UCSC Cancer Browser (Goldman et al., 2013).

2 RESULTS AND DISCUSSION

We developed TDM, a new method of RNA-seq data normalization intended for prediction using machine learning models built on microarray data and improved clustering. TDM performed well compared to quantile normalization and log_2 transformation on a range of data.

2.1 For Unsupervised Clustering, TDM is More Robust to Noise with Simulated Data

TDM outperformed quantile normalization on a clustering task using data simulating a matched set of 400 samples with both microarray and RNA-seq data. The data contained 4 simulated conditions and mimic the difference in dynamic range between microarrays and RNA-seq at 20 different levels of global noise (see Section 1).

Unsupervised clustering was performed using the PAM algorithm on the 400 samples with a microarray-like distribution. The accuracy of classification was assessed as the proportion of samples that were placed in a cluster in which the majority of samples matched their own class (Fig. 1). With no additional noise, all methods performed the same. However, as noise increased, the TDM transformation resulted in a consistently more accurate clustering than quantile normalization. Log_2 transformation resulted in unstable performance. At some noise levels it had much better classification, at others it did not, suggesting that this result would not scale to a larger, and more realistic dataset.

For additional insight into differences in clustering, we used principal coordinate analysis to visualize the similarity between samples in the data (Fig. 2). This was done at the 3% additional noise level (a point about midway between no noise and when the results start to level off in Fig. 1). The figure shows that TDM results in slightly better separation of the 4 clusters along the first 2 principal coordinates than the other two methods.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

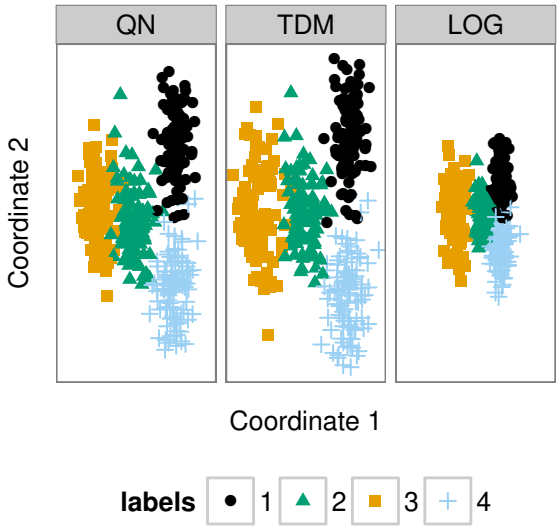


Figure 2. Principal coordinates of quantile (left), TDM (middle), and log_2 (right) normalized simulated data. The TDM normalized data result in slightly better separation along the two principal coordinates than the quantile and log_2 normalized data.

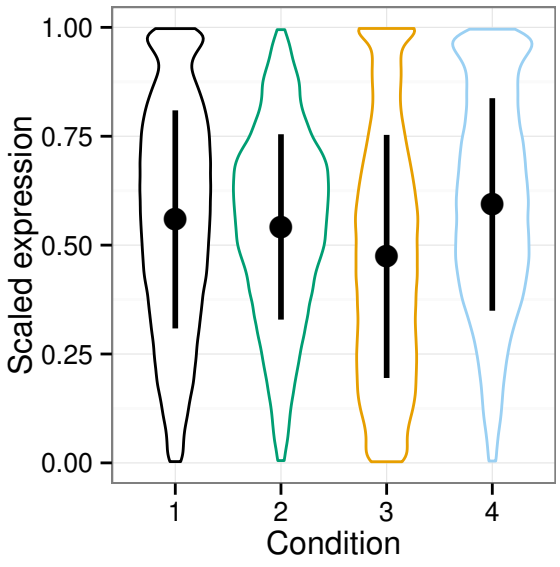


Figure 3. Violin plots of the simulated data distributions with standard deviation superimposed. These plots show the distribution of expression values for the samples with each particular condition. Within and between class variability is apparent in these data.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

2.2 Simulated Data Variability

Between vs. within class variability impacts the utility of data normalization methods, because if the within class variability outweighs the between class variability, it will be challenging to detect the signal of that condition in the data (Hicks and Irizarry, 2015). The distribution of expression values for each condition in the simulated data is shown with violin plots (Fig. 3), which display an appreciable level of variability both within and between classes. Quantro is a recently developed method for generating an F-score that represents the ratio of the within class variability to between class variability in the data (Hicks and Irizarry, 2015) and is available as a package for R. In particular, it provides guidance as to when quantile normalization should be applied so as to remove technical variation while minimizing the loss of biological signal. When the between class variation is low, then quantro indicates that quantile normalization should be applied. If the between class variation is much higher than the within class variation, then one must decide if it is likely to be biologically driven. If the variability is likely to be mostly technical, then quantile normalization may be effective. The simulated data provide an opportunity to assess the variability on data with tightly controlled conditions in order to better understand TDM's performance. We ran quantro on the simulated log_2 transformed RNA-seq data. The quantro score was approximately 2.01 which indicates that there is greater between class variability than within class variability and that quantile normalization may remove meaningful variability in the data if it is not mostly technical. As noise is added, the quantro score rises, eventually hitting 399.28 at 9.5% noise. This shows that as noise is added, the ratio of between class variability to within class variability rises. Of course, we know in this case that the difference in variability is technical, since we created it, but normally this information is not available, so quantro can provide useful guidance.

2.3 Evaluation of TDM for Supervised Model Construction using LASSO-Logistic Regression

For a supervised machine learning approach, we performed LASSO multinomial logistic regression to train models (on microarray datasets) for predicting tumor subtype in breast cancer and CIMP status in colon and rectal cancer, using the `glmnet` package (Friedman et al., 2010) in the R statistical environment. We then used the models to make predictions for RNA-seq datasets and the predictions were used to evaluate normalization techniques. We evaluated classification performance using the averaged values for each random seed for the total accuracy over all tumor subtypes/classes, balanced accuracy of each subtype/class, and Kappa statistic (classification rate after adjusting for those that could be expected by random chance).

2.3.1 Classification of breast cancer subtype on TCGA-only breast cancer dataset (Dataset 1)

TDM normalized data resulted in the highest mean total accuracy and Kappa on Dataset 1 (Fig. 4) across subtypes. The TDM normalized data had mean total accuracy of .63 and mean Kappa of .48. This was followed by quantile normalization with mean total accuracy of .57 and Kappa of .45. Log normalization had substantially lower mean total accuracy of .54 and Kappa of .45.

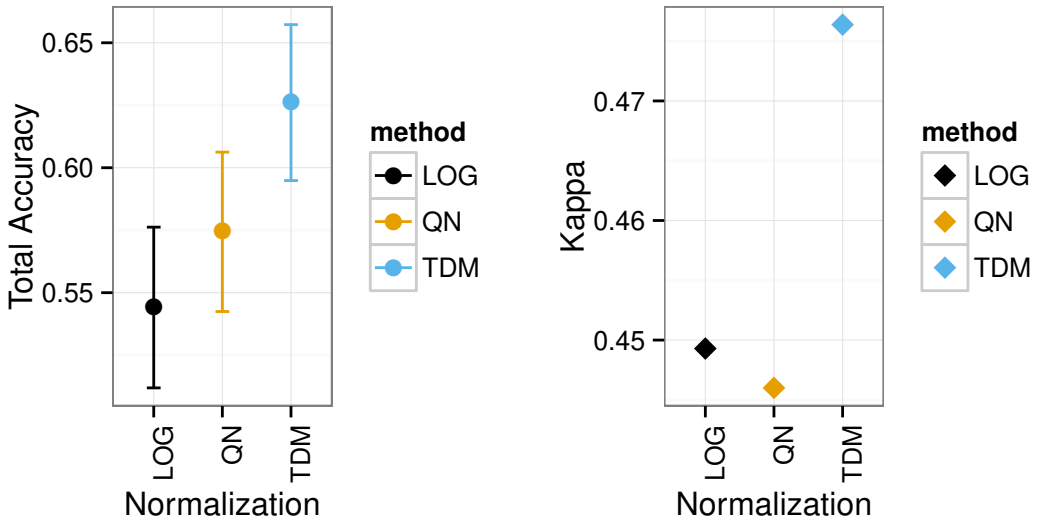
Within each subtype, there was considerable variability as to which normalization led to the best balanced accuracy on these data (Fig. S1). Quantile normalization resulted in best classification of Normal, but TDM and log_2 transformation were close. TDM resulted in best classification of LumA, and log_2 transformation resulted in best classification of Her2, LumB (although again TDM was close), and Basal. It is worth nothing that the distribution of samples for each subtype varies (Fig. 7).

2.3.2 Classification of CIMP status on TCGA-only colon/rectal cancer dataset (Dataset 2)

TDM normalized data resulted in the highest total accuracy and Kappa on Dataset 2 (Fig. 5). The TDM normalized data had mean total accuracy of .64, as well as mean Kappa of .36. This was very closely followed by quantile normalization with mean total accuracy of .62, although it had a Kappa of .29. Log normalization had somewhat lower mean total accuracy of .57, but its Kappa tied for best at .36.

Again, the normalization with the best balanced accuracy for specific tumor classes varied (Fig. S2). TDM resulted in best classification of CIMP (i.e. high positive CIMP status) and CIMPL (i.e. low positive CIMP status, although quantile normalization and log_2 were about the same). Log_2 transformation had the best classification for NCIMP (i.e. non-CIMP, although TDM was close) and Normal.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015



(a) Mean total accuracy for BRCA subtype classification across ten iterations. 95% confidence intervals shown. Dashed line represents the "no information rate" that could be achieved by always picking the most common class.

(b) Mean Kappa for BRCA subtype classification across ten iterations.

Figure 4. Results for Dataset 1: **(a)** TDM had the highest mean total accuracy on these data, followed by quantile normalization. **(b)** TDM had the highest mean Kappa on these data, followed closely by quantile normalization.

2.3.3 Classification of breast cancer subtype training on METABRIC and testing on TCGA (Dataset 3)

TDM and quantile normalization performed almost the same on Dataset 3 (Fig. 6), with mean total accuracies of .83 and .84 respectively, and both outperformed log_2 transformation, which had a mean total accuracy of 62. TDM and quantile normalization were more similar in classification to a separate dataset created from microarrays on the same samples, which had a mean total accuracy of .85, than they were to log_2 transformation of RNA-seq data. TDM and quantile normalization both had high Kappa scores on these data at .76 and .77 respectively. Log_2 transformation had a Kappa of .51 and the TCGA microarray data had a Kappa of .78.

For breast cancer subtypes (Fig. S3), in each case quantile normalization and TDM had better balanced accuracy than log_2 transformation (although it was close for Basal and LumB). The one subtype where the TCGA microarray performed substantially better was Her2.

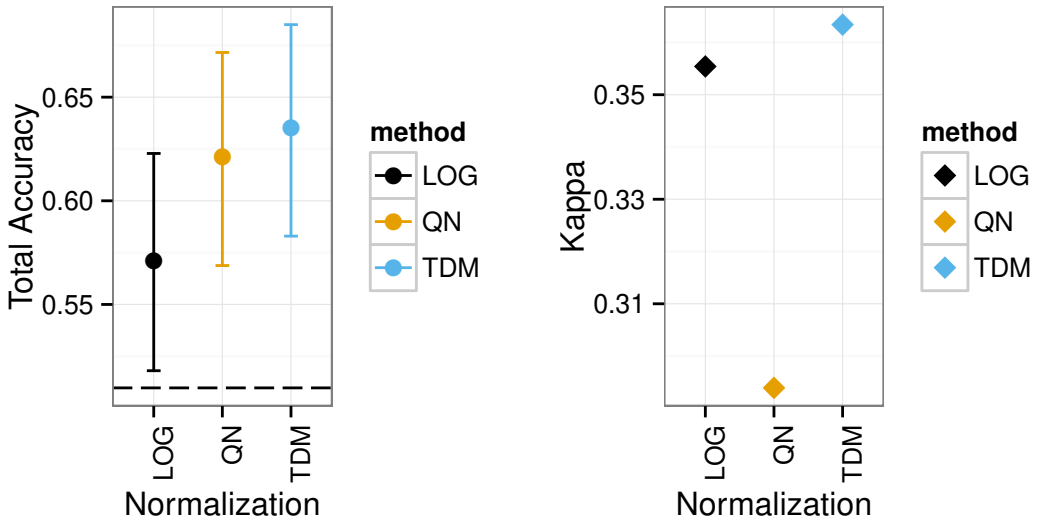
2.3.4 Summary of supervised machine learning applications

TDM resulted in the best performance overall on these datasets. For Dataset 1 and Dataset 2 it had the highest total accuracy and Kappa. For Dataset 3, quantile normalization had a very slightly higher total accuracy and Kappa than TDM, but only by about 1/2 of a percentage point and both were clearly better than log_2 transformation. For analyses where an independent microarray dataset was available, cross platform (microarray to RNA-seq) performance was comparable to within platform (microarray to microarray) performance for both quantile normalization and TDM.

2.4 Discussion

RNA-seq data transformed by the TDM algorithm outperformed those transformed by log_2 transformation or quantile normalization in most instances. The performance of quantile normalization could be sensitive to differing distributions of classes in training and test data. However, Fig. 7 shows that the distribution is roughly the same in each for all three biological datasets. Therefore, the difference in performance is probably attributable to noise. On the simulated data, TDM was consistently more robust against

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015



(a) Mean total accuracy for colon/rectal cancer CIMP classification across ten iterations. 95% confidence intervals shown. Dashed line represents the "no information rate" that could be achieved by always picking the most common class.

(b) Mean Kappa for colon/rectal cancer CIMP classification across ten iterations.

Figure 5. Results for Dataset 2: **(a)** TDM had the highest mean total accuracy, although it was only slightly better than quantile normalization. **(b)** TDM's Kappa was substantially higher than that achieved by log_2 transformation and both had Kappas higher than that achieved by quantile normalization.

noise, and these results support that assessment on biological data as well. Nevertheless, overall, quantile normalization performed only slightly worse than TDM. In particular, if the data are filtered to remove genes with low variance before training, our results support the use of either quantile normalization or TDM to obtain results with high accuracy. The implementation of such a step is dependent on the machine learning method used, and the goals of the study.

A factor in deciding to use quantile normalization will be the source of variance. Hicks et al. showed that when there is large variability across classes in the data and small within class variation that quantile normalization should not always be used (Hicks and Irizarry, 2015). At least some of the variance in these data may be attributable to the combination of colon and rectal cancer into a single dataset or due to difference in the distribution of subtypes and classes. In such a case, over-normalizing the data may also remove the signal. TDM provides an alternative: bring the values in the data more closely in line, while preserving inter-observation dependencies. This allows machine learning methods to better identify the signal that overcomes the noise of technical variability.

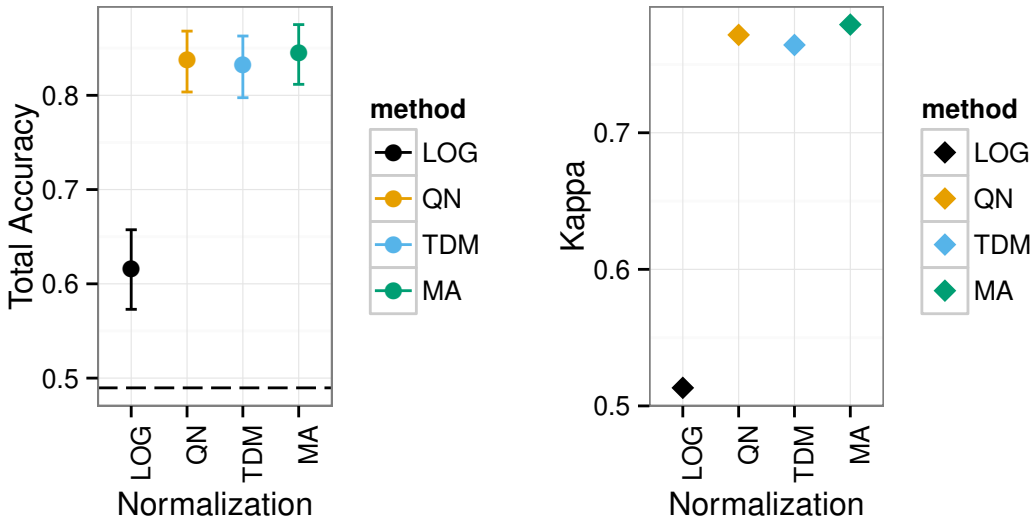
The results on Dataset 3, where both array and sequencing-based data were available, provide support for the use of the TDM algorithm for combining microarrays and RNA-seq in a single analysis. In this case, we had an additional microarray dataset measured on the same samples. TDM normalized data performed almost as well as an actual microarray dataset. This suggests that models built on data from one platform can be applied to another to generate meaningful predictions.

3 CONCLUSIONS

We developed TDM, a new method to normalize data so that models can be trained and evaluated without regard to platform. This will allow researchers to take advantage of the wealth of historical microarray data, including their own past experiments, as well as existing computationally derived models during the transition to next generation sequencing. We provide an R package for the transformation under the permissive open source BSD 3-clause license.

Our TDM algorithm successfully adjusts for the dataset shift that results from measurement on

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015



(a) Mean total accuracy for BRCA subtype classification across ten iterations using METABRIC microarray training data and TCGA RNA-seq test normalized data as well as TCGA microarray data for comparison (MA). 95% confidence intervals shown.

(b) Mean Kappa for BRCA subtype classification across ten iterations using METABRIC microarray training data and TCGA RNA-seq test normalized data as well as TCGA microarray data for comparison (MA).

Figure 6. Results for Dataset 3: **(a)** TDM and quantile normalization had the highest mean total accuracy for the normalized RNA-seq data when tested using a model trained on METABRIC. In fact, they were only slightly worse than actual microarray data from TCGA using the same samples, while log_2 transformation performed markedly worse. **(b)** TDM and quantile normalization achieved a high Kappa when tested using a model trained on METABRIC. They performed similarly to the TCGA microarray data (MA) that was assayed on the same samples.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

divergent platforms, such as that caused by the different dynamic ranges of microarrays and RNA-seq. TDM transforms the test data to have a similar distribution to the training data, while preserving most observation dependencies within those data. Because expression data are long-tailed, the compression of data near the end of the tail is expected to have a minimal impact for most machine learning methods. The consistent results with both unsupervised and supervised learning approaches on a variety of data support these conclusions and the broad utility of TDM.

4 DATA AVAILABILITY

TDM is available as an R package (Thompson and Greene, 2015a): https://github.com/greenelab/TDM

Code to apply TDM and alternative approaches to the datasets in this paper, analyze the results, and generate the figures is available in a parallel repository:
(Thompson and Greene, 2015b): https://github.com/greenelab/TDMresults

REFERENCES

Zeynep Kalender Atak, Valentina Gianfelici, Gert Hulselmans, Kim De Keersmaecker, Arun George Devasia, Ellen Geerdens, Nicole Mentens, Sabina Chiaretti, Kaat Durinck, Anne Uyttebroeck, et al. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in t-cell acute lymphoblastic leukemia. *PLoS genetics*, 9(12):e1003997, 2013. doi: 10.1371/journal.pgen.1003997.

Benjamin Milo Bolstad. *preprocessCore: A collection of pre-processing functions*, 2015. R package version 1.30.0.

B.M. Bolstad, R.a Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, (2):185–193, January . ISSN 1367-4803. doi: 10.1093/bioinformatics/19.2.185.

Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012. doi: 10.1038/nature10983.

Jaume Forés-Martos, Raimundo Cervera-Vidal, Enrique Chirivella, Alberto Ramos-Jarero, and Joan Climent. A genomic approach to study down syndrome and cancer inverse comorbidity: untangling the chromosome 21. *Frontiers in physiology*, 6, 2015. doi: 10.3389/fphys.2015.00010.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Paul Geeleher, Nancy J Cox, and R Stephanie Huang. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome biology*, (3):R47, January . ISSN 1465-6914. doi: 10.1186/gb-2014-15-3-r47.

Mary Goldman, Brian Craft, Teresa Swatloski, Kyle Ellrott, Melissa Cline, Mark Diekhans, Singer Ma, Chris Wilks, Josh Stuart, David Haussler, et al. The ucsc cancer genomics browser: update 2013. *Nucleic acids research*, 41(D1):D949–D954, 2013. doi: 10.1093/nar/gks1008.

Stephanie C Hicks and Rafael A Irizarry. quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome biology*, 16(1):117–117, 2015. doi: 10.1186/s13059-015-0679-0.

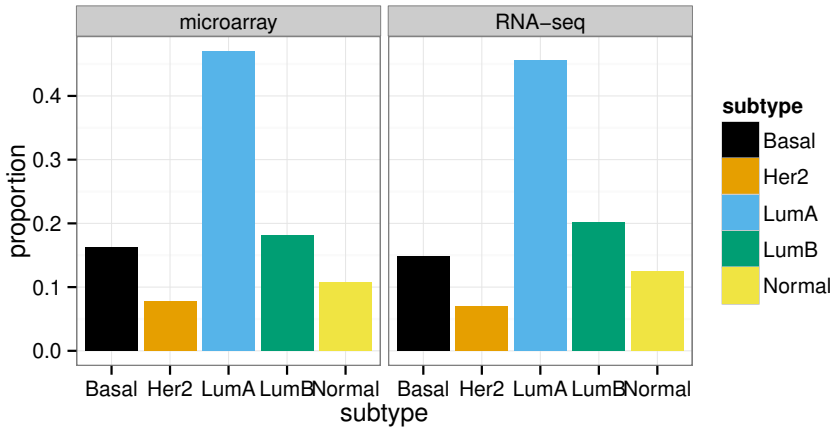
Curtis Huttenhower, Mark Schroeder, Maria D Chikina, and Olga G Troyanskaya. The Sleipnir library for computational functional genomics. *Bioinformatics (Oxford, England)*, (13):1559–61, July . ISSN 1367-4811. doi: 10.1093/bioinformatics/btn237.

Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, pages 68–125, 1990.

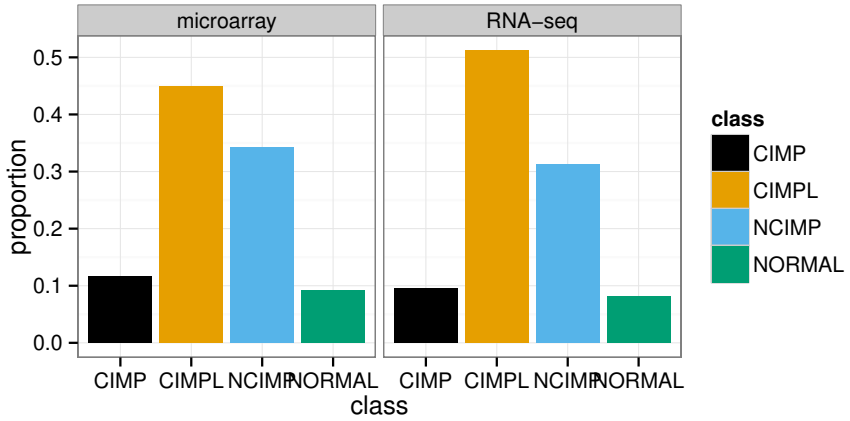
Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 2014. doi: 10.1016/j.csbj.2014.11.005.

Bin Li, Hyunjin Shin, Georgy Gulbekyan, Olga Pustovalova, Yuri Nikolsky, Andrew Hope, Marina Bessarabova, Matthew Schu, Elona Kolpakova-Hart, David Merberg, et al. Development of a drug-

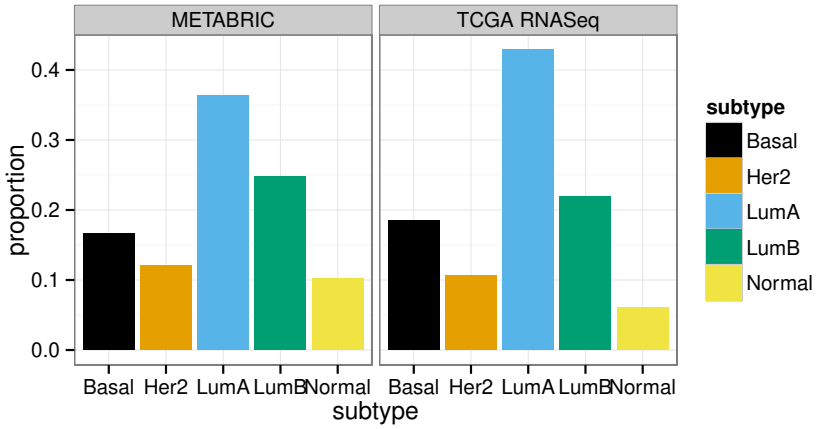
PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015



(a) Distribution of classes for the first biological dataset, using TCGA microarray of breast cancer biopsies for training and TCGA RNA-seq for testing.



(b) Distribution of classes for the second biological dataset, using TCGA microarray of colon/rectal cancer biopsies for training and TCGA RNA-seq for testing.



(c) Distribution of classes for the third biological dataset, using METABRIC microarray of breast cancer biopsies for training and TCGA RNA-seq for testing.

Figure 7. The distribution of classes in the data is roughly the same between each training and testing set.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PloS one*, 10(6), 2015. doi: 10.1371/journal.pone.0130700.

Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, and Hai Zhang. Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. *BMC bioinformatics*, (1):198, January . ISSN 1471-2105. doi: 10.1186/1471-2105-14-198.

Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2015. R package version 2.0.1 — For new features, see the 'Changelog' file (in the package source).

John H Malone and Brian Oliver. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, page 34, January . ISSN 1741-7007. doi: 10.1186/1741-7007-9-34.

Marie Mooney, Jeffrey Bond, Noel Monks, Emily Eugster, David Cherba, Pamela Berlinski, Steve Kamerling, Keith Marotti, Heather Simpson, Tony Rusk, Waibhav Tembe, Christophe Legendre, Hollie Benson, Winnie Liang, and Craig Paul Webb. Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PloS one*, (4):e61088, January . ISSN 1932-6203. doi: 10.1371/journal.pone.0061088.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, (1):521–530, January . ISSN 00313203. doi: 10.1016/j.patcog.2011.06.019.

Nadine Norton, Zhifu Sun, Yan W Asmann, Daniel J Serie, Brian M Necela, Aditya Bhagwate, Jin Jen, Bruce W Eckloff, Krishna R Kalari, Kevin J Thompson, et al. Gene expression, single nucleotide variant and fusion transcript discovery in archival material from breast tumors. *PLOS One*, 2013. doi: 10.1371/journal.pone.0081925.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Jon Ison, Maria Keays, et al. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic acids research*, (Database issue):D987–90, January . ISSN 1362-4962. doi: 10.1093/nar/gks1174.

Francisco Sánchez-Vega, Valer Gotea, Gennady Margolin, and Laura Elnitski. Pan-cancer stratification of solid human epithelial tumors and cancer cell lines reveals commonalities and tissue-specific features of the cpg island methylator phenotype. *Epigenetics & chromatin*, 8(1):14, 2015. doi: 10.1186/s13072-015-0007-7.

Jia Tan, Matthew Ung, Chao Cheng, and Casey Greene. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. In *Proceedings of PSB 2015*. Pacific Symposium on Biocomputing, 2015.

Matthias Templ, Alexander Kowarik, and Bernhard Meindl. *sdcMicro: Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation*, 2015. URL `http://CRAN.R-project.org/package=sdcMicro`. R package version 4.5.0.

The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature*, (7418): 61–70, October . ISSN 1476-4687. doi: 10.1038/nature11412.

Jeffrey A. Thompson and Casey S. Greene. TDM: Submitted with preprint and manuscript, October 2015a.

Jeffrey A. Thompson and Casey S. Greene. Training Distribution Matching (TDM) Evaluation and Results, October 2015b.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, (6):520–525, June . ISSN 1367-4803. doi: 10.1093/bioinformatics/17.6.520.

Karolis Uziela and Antti Honkela. Probe region expression estimation for rna-seq data for improved microarray comparability. *PLOS One*, 2015. doi: 10.1371/journal.pone.0126545.

Tim Van den Bulcke, Koenraad Van Leemput, Bart Naudts, Piet van Remortel, Hongwu Ma, Alain Verschoren, Bart De Moor, and Kathleen Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*, 7(1):43, 2006. doi: 10.1186/1471-2105-7-43.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015

448 Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu,
449 Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, et al. The concordance between rna-seq and
450 microarray data depends on chemical treatment and transcript abundance. *Nature biotechnology*, 32(9):
451 926–932, 2014. doi: 10.1038/nbt.3001.
452 Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq : a revolutionary tool for transcriptomics.
453 *Nature Reviews Genetics*, 10(1):57–63, 2010. doi: 10.1038/nrg2484.RNA-Seq.
454 Iris H Wei, Yang Shi, Hui Jiang, Chandan Kumar-Sinha, and Arul M Chinnaiyan. Rna-seq accurately
455 identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia*, 16(11):918–927, 2014.
456 doi: 10.1016/j.neo.2014.09.007.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1460v1 | CC-BY 4.0 Open Access | rec: 30 Oct 2015, publ: 30 Oct 2015