BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 956–961
doi:10.1093/bioinformatics/btt574

Gene expression

Advance Access publication October 11, 2013

PLIDA: cross-platform gene expression normalization using perturbed topic models

Amit G. Deshwar[1,*] and Quaid Morris[1,2,3,4]

[1]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, [2]Department of Molecular Genetics, [3]Department of Computer Science and [4]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 1A1, Canada

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: Gene expression data are currently collected on a wide range of platforms. Differences between platforms make it challenging to combine and compare data collected on different platforms. We propose a new method of cross-platform normalization that uses topic models to summarize the expression patterns in each dataset before normalizing the topics learned from each dataset using per-gene multiplicative weights.

Results: This method allows for cross-platform normalization even when samples profiled on different platforms have systematic differences, allows the simultaneous normalization of data from an arbitrary number of platforms and, after suitable training, allows for online normalization of expression data collected individually or in small batches. In addition, our method outperforms existing state-of-the-art platform normalization tools.

Availability and implementation: MATLAB code is available at http://morrislab.med.utoronto.ca/plida/.

Contact: Amit.Deshwar@utoronto.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 30, 2012; revised on June 19, 2013; accepted on September 27, 2013

1 INTRODUCTION

Microarray platforms differ from one another in their manufacture, labeling methods, hybridization procedures, number of probes per gene, probe length and probe sequences. Although these characteristics can affect the dynamic range, precision and accuracy of the resulting gene expression measurements, the major modern probe-based microarrays (e.g. Affymetrix, Agilent, Applied Biosystems (ABI) and Illumina) provide data of similar quality, so long as they are normalized appropriately (Shi *et al.*, 2006).

However, even microarrays from the same provider can differ in the probes used to target specific genes, and these differences can lead to systematic differences in the relationship between their target abundance and the measured probe intensity (Shi *et al.*, 2006). The presence of these platform effects is well known and often acknowledged by the admission that gene expression levels are measured in arbitrary units. However, because most comparisons in gene expression levels are done on a per-gene basis among samples from the same platform, these platform-specific biases have limited impact and can be safely ignored. However, quantitative comparisons between gene expression datasets from different microarray platforms benefit from correcting these biases (Rudy and Valafar, 2011). Given the significant cost of collecting and analyzing human tissue, there are significant benefits to attempting to combine gene expression measurements from multiple studies to increase statistical power or provide a broader overview of disease.

The most straightforward method of cross-platform normalization is to standardize the expression of each gene to have the same center (using the median or the mean) and the same variance on the different platforms using a gene-specific affine transform. This method has the advantage of being fast, easy and having no parameters to tune, but it does not work as well as more complex normalization methods on supporting, for example, estrogen receptor (ER) status prediction by integrating data across platforms (Shabalin *et al.*, 2008). Of the many different cross-platform normalization methods that were compared in a recent review by Rudy and Valafar (2011), distance-weighted discrimination (DWD) and cross-platform normalization were identified as superior choices.

Cross-platform normalization (Shabalin *et al.*, 2008) consists of a two-step procedure. In the first step, k-means clustering is used to find blocks of similar genes and samples across the platforms. Then, within each block, cross-platform normalization fits a combination of additive and multiplicative weights to normalize the data between platforms within this block. This process is done multiple times and the results from each run are averaged together to account for various possible clustering patterns. Their article reported the ability of cross-platform normalization to significantly increase the ability to use data collected on one platform as training data to predict ER status in breast cancer samples collected on a second platform.

DWD (Benito *et al.*, 2004) consists of finding the direction that best separates the data from the two platforms in log-vector space and then translating data from each platform along that direction until the data from each platform overlaps. Implicit in DWD's formulation is the assumption that gene expression data are log-normally distributed [c.f. Rocke and Durbin (2001)] because it is based on linear discriminant analysis, and that platform differences are the only systematic differences between the two datasets are separated.

*To whom correspondence should be addressed

Despite this existing work, there are still unsolved problems in cross-platform normalization. One is normalizing data collected on more than two platforms. Although it is possible to chain cross-platform normalization steps together, it is not clear what effect multiple normalization steps will have on the data and what the preferred chaining strategy (tree-based or linear) should be. Another unaddressed problem in cross-platform normalization is 'online use'; in other words, normalizing data using a normalization function learned from a previously collected dataset. Online normalization would be useful, for example, for applying a gene signature derived from another platform to new gene expression profiles where the cross-platform normalization was previously defined using a calibration dataset.

We are also interested in addressing use cases where cross-platform normalization is applied among heterogeneous datasets that differ in their sample composition. Many methods (including DWD but not cross-platform normalization) assume that systematic differences between datasets are because of platform effects and can 'over-normalize' by removing all systematic expression differences between the two datasets, including those due to differences in the samples represented. For example, breast cancer datasets can differ in their expression subtype composition and in these cases, over-normalization would interfere with gene signature construction on integrated datasets. Over-normalization can also interfere with the use of cross-platform normalization in computational purification in which expression profiles of normal tissue are used to try to remove the impact of normal tissue contamination in tumor expression profiles (Clarke *et al*., 2010; Nicolau *et al*., 2007; Quon and Morris, 2009; Quon *et al*., 2013).

One way of dealing with heterogeneity between datasets is to assume that the variation is found in a low-dimensional subspace of the dataspace, using dimensionality reduction techniques such as Principal Components Analysis (PCA). Dimensionality reduction is often used when analyzing gene expression data from multiple platforms. For example, Wang *et al*. (2011) use dimensionality reduction to summarize gene expression measurements on multiple platforms of the same samples, but they do not attempt to make measurements on different platforms comparable. However, because gene expression analysis measures the quantity of messenger RNA transcripts for a particular gene, it is ultimately measuring a non-negative quantity. This makes topic models, such as the latent Dirichlet allocation (LDA) (Blei *et al*., 2003), particularly appropriate. LDA models count observations as coming from a convex combination of multinomial distributions and has been used previously for analysis of gene expression data (Gerber *et al*., 2007; Gevaert *et al*., 2006).

2 APPROACH

Our method is based on the hypothesis that, in the absence of platform-induced differences, the gene expression profiles from different platforms should be able to be summarized using the same topics. Based on this hypothesis, we propose to carry out cross-platform normalization by simultaneously learning a topic model decomposition that describes the data from all the platforms being normalized and per-platform adjustment weights that modify the topics to account for platform biases. We call our method Platform Independent LDA, or PLIDA.

This method addresses several issues with cross-platform normalization not yet addressed with existing methods. This method naturally handles the case of normalizing an arbitrary number of platforms simultaneously, using all the data to learn the underlying structure of the data. Using our method, one could learn the platform-specific adjustment weights using archival data and then apply those weights to data collected in the future, enabling online use and simplifying the transition of previously defined clinical biomarkers to new platforms. Finally, our method explicitly models differences between datasets and does not strive to eliminate all differences, just those ones not explainable by a topic model decomposition.

PLIDA assumes a one-to-one mapping between genes (or transcripts) represented on all of the platforms, and data from platforms with multiple probes per gene should be summarized at the gene level before applying PLIDA. PLIDA also assumes that the gene expression measurements are made in the linear range of the platform being used, a reasonable approximation for most genes on modern expression platforms (Shen-Orr *et al*., 2010).

3 METHODS

3.1 Generative model and inference

Our generative model for gene expression measurements from multiple platforms is similar to that of LDA (Blei *et al*., 2003), except that each platform has a set of per-gene multiplicative scaling factors (called 'adjustment factors') that modify the topics used to model samples from that platform. Formally, our generative model can be described as follows:

For each observed sample d from $\{1, \ldots, M\}$:

(1) Select a platform p from $\{1, \ldots, P\}$

(2) For each observed transcript t_n in the sample:

(a) Select a topic z from $\{1, \ldots, K\}$ using a sample-specific discrete distribution over topics θ_d.

(b) Select a transcript. The transcript will be selected from a distribution specific to the topic selected in step 2 and the platform selected in step 1. The probability of drawing transcript g, given z and p is as follows:

$$P(t_n == g) = f_g^p * \pi_g^z \Big/ \sum_{g'} f_{g'}^p * \pi_{g'}^z$$

In this equation, f_g^p is a platform- and gene-specific adjustment factor that modifies the probability of drawing that transcript relative to the 'base' topic distribution π^z. The adjustment factor itself is drawn from a gamma distribution. These probabilistic relationships and the priors associated with the different variables are summarized below:

$$\theta_d \sim Dirichlet(\alpha)$$

$$z \sim Discrete(\theta_d)$$

$$\pi^z \sim Dirichlet(\beta)$$

$$f_g^p \sim Gamma(a, b)$$

The plate diagram describing this model can be found in Figure 1.

The drawing of the perturbation factor f_g^p from a gamma distribution limits the range of perturbation factors to be greater than zero. Furthermore, we set $b = 1/a$, so the mean of the Gamma prior is

$a * b = 1$. Fitting this model to the gene expression data required the discretization of the observations, which were rounded to the nearest integer. To estimate the parameters of the model, we run 50 iterations of an optimization procedure that maximizes the posterior log likelihood using a block conjugate gradient descent approach. In each iteration, the Polak–Ribiere conjugate gradient descent (Hestenes and Stiefel, 1952) is used to optimize the following sets of variables: $\{\pi^1\}, \ldots, \{\pi^K\}, \{\theta_1\}, \ldots, \{\theta_M\}, \{\alpha\}, \{f^1\}, \ldots, \{f^P\}$

Variables within a set are optimized simultaneously, whereas each set is optimized sequentially once per iteration.

Although this method has many hyperparameters, for all experiments reported in this manuscript the number of topics was set to 2, and a and b were set to 0.25 and $4 = 1/0.25$, respectively. The value of a should be decreased (or increased) if stronger (or weaker) perturbations are expected on average. Although the user can change the number of topics (K), we recommend setting the number of topics to 2, as we found that this provided the ability to model platform-independent systematic differences among datasets (e.g. due to differences in sample composition) while preventing the model from explaining platform-dependent differences using platform-specific topics (data not shown).

3.2 Gene expression datasets used

3.2.1 Prostate datasets The first prostate tumor dataset, from Wang *et al.* (2010) consists of 140 prostate tumor samples collected on the Affymetrix U133A platform, available as Gene Expression Omnibus (GEO) ascension GSE8218. The second prostate dataset comes from the same manuscript, but uses the Affymetrix U133 Plus2 platform and consists of 45 normal prostate samples, available as GEO ascension GSE17951. Both prostate datasets were normalized within each dataset using Robust Multi-array Average (RMA) (Irizarry *et al.*, 2003) and mapped to Entrez gene IDs using a custom Chip Definition File (CDF) (Dai *et al.*, 2005).

3.2.2 Breast cancer datasets Dataset 1 comes from Hu *et al.* (2006) and is collected on the Agilent Human 1A (V2) platform and consists of 69 breast tumor samples, 24 of which are ER positive (GEO ascension GSE1992). This dataset was Lowess normalized Cleveland (1979) and mapped to Entrez IDs by the original authors. Dataset 2 (Desmedt *et al.*, 2007) uses the Affymetrix U133A platform and consists of 198 breast cancer samples, 134 are ER positive (GSE 7390). Dataset 3 comes from the MicroArray Quality Control (MAQC) II project (Shi *et al.*, 2010) and consists of 278 breast cancer samples with 164 ER positive cases, using the Affymetrix U133A platform and available as GSE20194. Breast datasets 2 and 3 were normalized within each dataset using RMA (Irizarry *et al.*, 2003) and mapped to Entrez gene IDs using a custom CDF file (Dai *et al.*, 2005).

For the breast and prostate datasets, only the genes measured on all the compared platforms were used. The Entrez gene IDs provided by the original authors (breast dataset 1) or provided by the custom CDF file (prostate datasets, breast datasets 2 and 3) were used to find the shared genes.

3.2.3 MAQC dataset This dataset comes from the MAQC project (Shi *et al.*, 2006) and was used by Rudy and Valafar (2011). It consists of 299 samples composed of technical replicates of four different mixtures of Stratagene Universal Human Reference RNA and Ambion Human Brain Reference RNA collected on four different platforms. In all, 120 samples were collected on the Affymetrix U133 Plus2 platform, 59 samples on the Illumina Sentrix Human-6 Expression Beadchip, 60 samples on the ABI Human Genome Survey Microarray and 60 samples on the Agilent Whole Human Genome Oligo Micoarray G4131A. These data are available already normalized as the R package CONORData (Rudy and Valafar, 2012).

4 RESULTS

4.1 Assumption testing

PLIDA assumes that gene expression data can be decomposed into a small number of topics and that biases between platforms can be corrected with a per-gene adjustment factor. To test these assumptions we carried out two experiments. In the first, we used prostate dataset 1, the prostate tumor dataset. Prostate tumor samples are composed of varying proportions of tumor and normal tissue, so this dataset is not homogeneous as would be expected from a sample from one tissue type. This dataset was randomly partitioned into two halves, representing two 'different' platforms. Because there is no difference in platform characteristics, we expect that the platform-specific factors learned should all be close to one. Figure 2 shows a histogram of the platform-specific factors learned for the pseudo-platform. In the log2 domain, most values are tightly clustered around 0 (i.e. no adjustment).

As another experiment, we verified that our method was capable of learning systematic differences between two datasets. Taking the same dataset as used the previous experiment, we perturbed the expression values in one half of the dataset. Perturbations were drawn randomly from a log-uniform distribution between one-third and three. After learning, we compared the known perturbations with the learned perturbations, shown in Figure 3. The concordance between the learned perturbations and the known values was good, with a Pearson correlation of

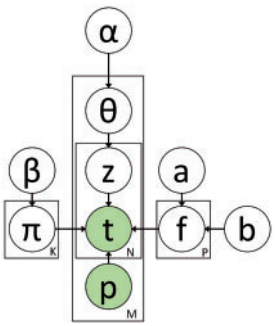

Fig. 1. Graphical model for PLIDA

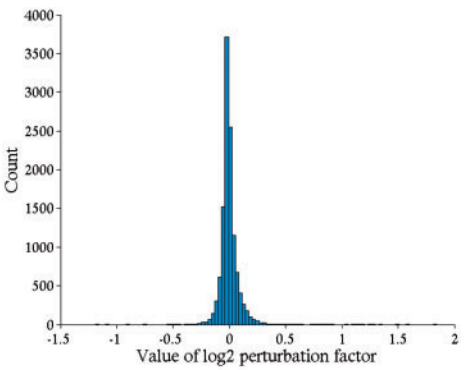

Fig. 2. Distribution of log platform adjustment factors when datasets from the same platform are normalized

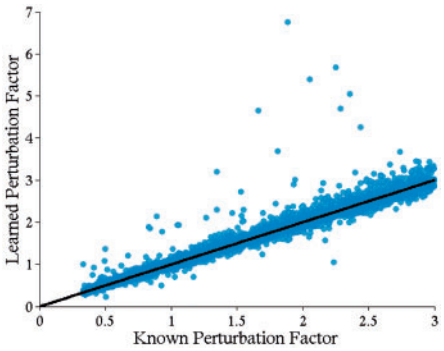

Fig. 3. Scatter plot between known and learned perturbation factors. Black line shows $y = x$ line

0.97 and most values lying close to the $y = x$ line. The ability to recover known perturbations in real gene expression data, even though the perturbations were drawn from a different distribution than the model uses, increases our confidence in PLIDA.

We repeated this experiment for 3- and 4-way splits, each time recovering the known perturbations with high accuracy. For the 3-platform case, the Pearson correlations between the known and recovered perturbations were 0.97 and 0.98. For the 4-platform case, the Pearson correlations between the known and recovered perturbations were 0.97, 0.94 and 0.95 (See Supplementary Figs. 1 and 2).

4.2 MAQC

To demonstrate the ability of PLIDA to remove systematic biases in microarray data, we plot all 299 datapoints from the MAQC dataset in the first two principal component spaces before and after normalization by PLIDA (Fig. 4). Before normalization, data collected on each platform clustered with others from the same platform, while this tendency is greatly reduced after normalization. This reduction in differences across platforms is confirmed by median concordance between the platforms that ranged from 0.45 to 0.76 before normalization but were above 0.98 after normalization. We also plot the topic loadings (θ) found, showing that samples from different platforms are well dispersed, despite the presence of some intra-platform clustering.

4.3 ISOpure testing

In this section, we test the utility of gene expression normalization for use with ISOpure, a computational method for purifying tumor expression profiles and inferring tumor cellularity (Quon *et al.*, 2013). For this experiment, we used both prostate datasets described in the Methods section. For each prostate tumor sample, the sample was examined by a pathologist who provided an estimate of the cancerous composition of the sample. The aim of the experiment was to normalize the dataset of 45 normal samples so it could be used to generate accurate predictions of the cancerous composition of the tumor samples using ISOpure.

We normalized the normal prostate dataset with the tumor dataset using either PLIDA, DWD or cross-platform normalization. We then ran ISOpure using the original and normalized tumor and normal tissue datasets. We then compared the

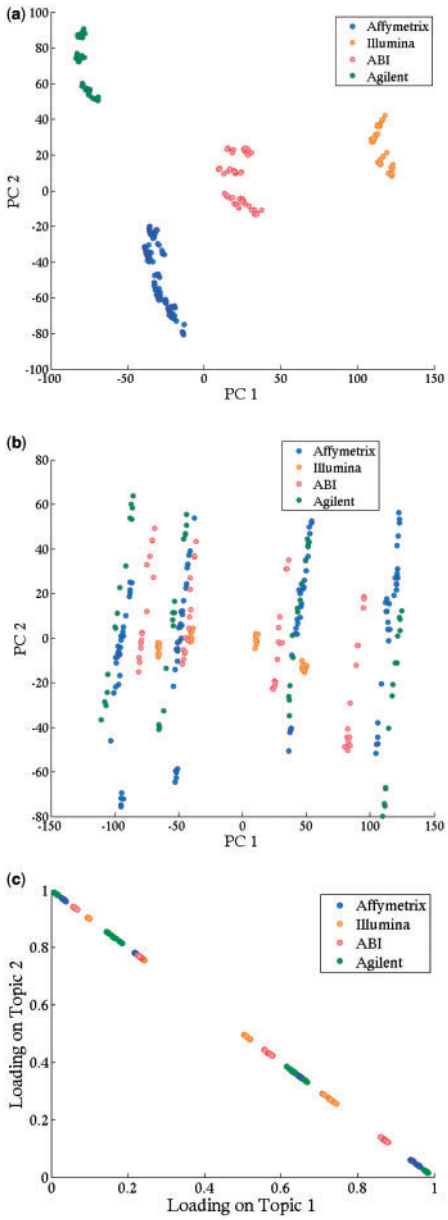

Fig. 4. Scatter plots showing the first two principal components of the MAQC data, before (**a**) and after (**b**) normalization, and topic loadings learned by PLIDA (**c**)

cancerous composition estimates from ISOpure for each dataset version (original, PLIDA, cross-platform normalization and DWD) with the pathologist's estimates. Without any normalization, the Spearman correlation between ISOpure's estimates and the pathologist was 0.2153. Normalizing the data using DWD or cross-platform normalization resulted in noticeably worse performance than the original case, with correlations of 0.0059 and 0.1046, respectively. In contrast, using PLIDA greatly increased the correlation between ISOpure estimates and the pathologist's, with a correlation of 0.3502 (see Fig. 5). The correlation found using PLIDA normalized data was significantly greater than that found using cross-platform normalization and

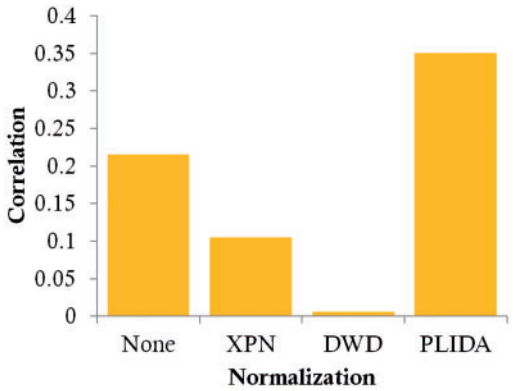

Fig. 5. Correlation between ISOpure's cancerous composition estimates and pathologist

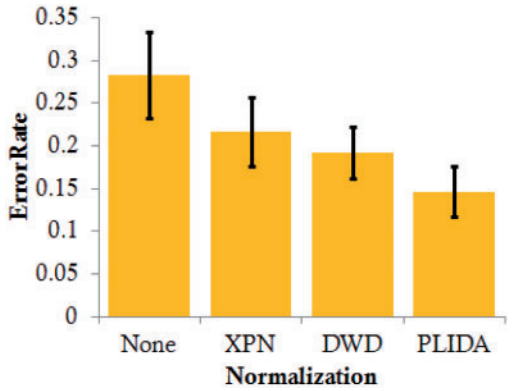

Fig. 7. Error rate for classifiers trained on breast dataset 1 and tested on breast dataset 2

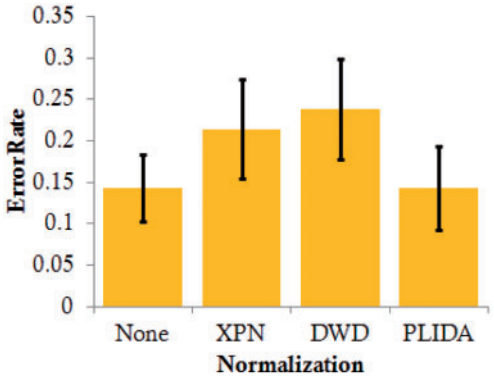

Fig. 6. Error rate for classifiers trained on breast dataset 2 and tested on breast dataset 1

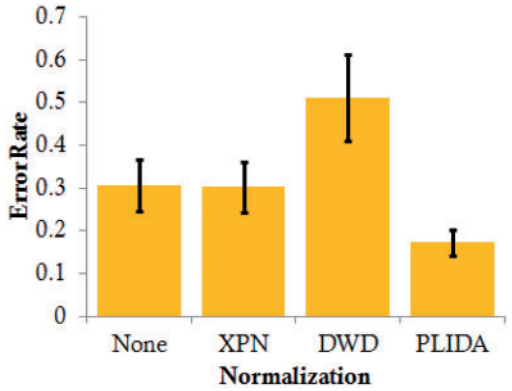

Fig. 8. Error rate for classifiers trained on breast dataset 1 (normalized with breast dataset 2) and tested on breast dataset

DWD ($P<0.05$, Steiger's Z-test). In examining the topic loadings of the optimized PLIDA model, the normal prostate dataset had almost all of its mass (98%) on one of the topics, whereas the cancerous dataset had 33% of its mass on the same topic. This pattern is what we would expect, given that the tumor samples are composed of some normal tissue along with cancerous tissue.

4.4 ER status classification

In this section, we test the ability of using gene expression normalization to train a classifier to determine (ER) status from breast cancer gene expression data. This task is similar to one used in Shabalin *et al.* (2008) to validate their cross-platform normalization algorithm. We trained elastic-net-regularized logistic regression classifiers (Friedman *et al.*, 2010) using breast dataset 1 and then measured the classification error rate on breast dataset 2 before reversing the procedure. An identical procedure was carried out after normalizing data from the two datasets using PLIDA, DWD and cross-platform normalization. Classifier hyperparameters were set by an inner cross-validation loop on the training dataset. For training on breast dataset 2 and testing on breast dataset 1, accuracy was highest using the original expression data and with PLIDA-normalized data, with the other two normalization methods lowering the accuracy, as shown in Figure 6.

For training on breast dataset 1 and then testing on breast dataset 2, the results were more variable. Using the original datasets resulted in an error rate of 0.2828, much higher than the error rates obtained for the reverse case. All three normalization methods resulted in increased classifier performance, but PLIDA performed the best, as shown in Figure 7. Classifiers trained using PLIDA, cross-platform normalization and DWD had error rates of 0.1465, 0.1919 and 0.2163, respectively.

4.5 Online use

In this section, we demonstrate the ability of our method to learn adjustment factors using archival data, than apply these factors to normalize platform differences on data not available at training time. Again, the task was ER status prediction, and we normalized Breast Dataset 1 and Breast Dataset 2. We then trained an elastic-net logistic regression classifier to predict ER using the normalized Breast Dataset 1 and used this classifier to predict ER status of patients in the third Breast Dataset after applying the weights learned previously. With DWD and cross-platform normalization, the first and second breast datasets were normalized together and again only the data from the first dataset used for training. In this scenario, classifiers trained using PLIDA-normalized data performed much better than the original data

or the other normalization methods. Using the original data resulted in an error rate of 0.3058, similar to the error rate when tested against breast dataset 2. cross-platform normalization-normalized data lead to a similar error rate as the unnormalized data (0.3021), whereas DWD-normalized data did worse, with an error rate of 0.5111. In comparison, PLIDA was able to bring down the error rate to 0.1727, as shown in Figure 8.

5 DISCUSSION

Gene expression data provide an inexpensive and high-resolution snapshot of cellular state, but comparison and use of gene expression profiles collected on different platforms and technologies remain difficult. Methods of cross-platform normalization have already been developed but have limitations that make them unsuitable for some desired uses of gene expression data. We developed PLIDA to provide a cross-platform normalization tool to overcome these limitations, allowing the greater use of platform-varied gene expression data.

Specifically PLIDA is capable of normalizing gene expression data collected on different platforms even when there are systematic differences between the samples on each platform. It does this by finding a low-dimensional topic model representation of the data, and then normalizing this topic model across platforms. This allows for the integration of more datasets than cross-platform normalization methods that assume that all differences between experiments are solely due to the platform. PLIDA is also capable of online use, by learning platform adjustment weights using existing data and then applying these weights to data collected in the future. This allows data collected on different platforms to be used to train classifiers for data that have not been observed at the time of normalization, as when a biomarker signature is used in the clinic. Finally, PLIDA is capable of normalizing an arbitrary number of platforms simultaneously.

Funding: National Science and Engineering Research Council (NSERC) operating grant (to Q.M.); NSERC Julie Payette post-graduate fellowship (to A.G.D.).

Conflict of Interest: none declared.

REFERENCES

Benito,M. *et al.* (2004) Adjustment of systematic microarray data biases. *Bioinformatics*, **20**, 105–114.

Blei,D. *et al.* (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.

Clarke,J. *et al.* (2010) Statistical expression deconvolution from mixed tissue samples. *Bioinformatics*, **26**, 1043–1049.

Cleveland,W. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.*, **33**, e175.

Desmedt,C. *et al.* (2007) Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clin. Cancer Res.*, **13**, 3207–3214.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gerber,G.K. *et al.* (2007) Automated discovery of functional generality of human gene expression programs. *PLoS Comput. Biol.*, **3**, e148.

Gevaert,O. *et al.* (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, **22**, e184–e190.

Hestenes,M. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear systems. *J. Res. Natl Bur. Stand.*, **49**, 409–436.

Hu,Z. *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.

Irizarry,R. *et al.* (2003) Summaries of affymetrix genechip probe level data. *Nucleic Acids Res.*, **31**, e15.

Nicolau,M. *et al.* (2007) Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics*, **23**, 957–965.

Quon,G. and Morris,Q. (2009) Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, **25**, 2882–2889.

Quon,G. *et al.* (2013) Computational purification of individual tumor gene expression profiles. *Genome Med.*, **5**, 1–20.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, **8**, 557–569.

Rudy,J. and Valafar,F. (2011) Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinformatics*, **12**, 467.

Rudy,J. and Valafar,F. (2012) *CONORData: CONORData*. R package version 1.0.2, http://cran.at.r-project.org/web/packages/CONORData/index.html.

Shabalin,A. *et al.* (2008) Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, **24**, 1154–1160.

Shen-Orr,S. *et al.* (2010) Cell type–specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

Shi,L. *et al.* (2006) The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.*, **24**, 1151–1161.

Shi,L. *et al.* (2010) The Microarray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Wang,X.V. *et al.* (2011) Unifying gene expression measures from multiple platforms using factor analysis. *PLoS One*, **6**, e17691.

Wang,Y. *et al.* (2010) *In silico* estimates of tissue components in surgical samples based on expression profiling data. *Cancer Res.*, **70**, 6448–6455.