Bioinformatics, 31(14), 2015, 2318–2323
doi: 10.1093/bioinformatics/btv157
Advance Access Publication Date: 18 March 2015
Original paper

OXFORD

Gene expression

Test set bias affects reproducibility of gene signatures

Prasad Patil[1], Pierre-Olivier Bachant-Winner[2], Benjamin Haibe-Kains[3,4,*] and Jeffrey T. Leek[1,*]

[1]Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, MD, USA, [2]Institut de Recherches Cliniques de Montréal, Montreal, Quebec H2W 1R7, Canada, [3]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario M5G 1L7, Canada and [4]Department of Medical Biophysics, University of Toronto, Toronto, Ontario M5G 1L7, Canada

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on October 22, 2014; revised on February 13, 2015; accepted on March 16, 2015

Abstract

Motivation: Prior to applying genomic predictors to clinical samples, the genomic data must be properly normalized to ensure that the test set data are comparable to the data upon which the predictor was trained. The most effective normalization methods depend on data from multiple patients. From a biomedical perspective, this implies that predictions for a single patient may change depending on which other patient samples they are normalized with. This test set bias will occur when any cross-sample normalization is used before clinical prediction.

Results: We demonstrate that results from existing gene signatures which rely on normalizing test data may be irreproducible when the patient population changes composition or size using a set of curated, publicly available breast cancer microarray experiments. As an alternative, we examine the use of gene signatures that rely on ranks from the data and show why signatures using rank-based features can avoid test set bias while maintaining highly accurate classification, even across platforms.

Availability and implementation: The code, data and instructions necessary to reproduce our entire analysis is available at https://github.com/prpatil/testsetbias.

Contact: jtleek@gmail.com or bhaibeka@uhnresearch.ca

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

1 Introduction

One of the most common barriers to the development and translation of genomic signatures is cross-sample variation in technology, normalization and laboratories (Majewski and Bernards, 2011). Technology, batch and sampling artifacts have been responsible for the failure of genomic signatures (Baggerly *et al.*, 2005; Petricoin *et al.*, 2002), irreproducibility of genomic results (Michiels *et al.*, 2005) and retraction of papers reporting genomic signatures (Sebastiani *et al.*, 2010). Even highly successful signatures such as Mammaprint (van't Veer *et al.*, 2002) have required platform-specific retraining before they could be translated to clinical use (Glas *et al.*, 2006). An under-appreciated source of bias in genomic signatures is test set bias (Lusa *et al.*, 2007). Test set bias occurs when the predictions for any single patient depend on the data for other patients in the test set. For example, suppose that the gene expression data for a single patient is normalized by subtracting the mean expression and dividing by the standard deviation of the expression across all patients in the test set. Then the normalized value for any specific gene for that patient depends on the values for all the patients they are normalized with. The result is that a patient may get two different predictions using the same data and the same prediction algorithm, depending on the other patients used to normalize the test set data (Fig. 1).

There are many scenarios under which a patient's classification ought to change: if new information updates or alters the prediction algorithm or if the raw, biological patient data itself changes.

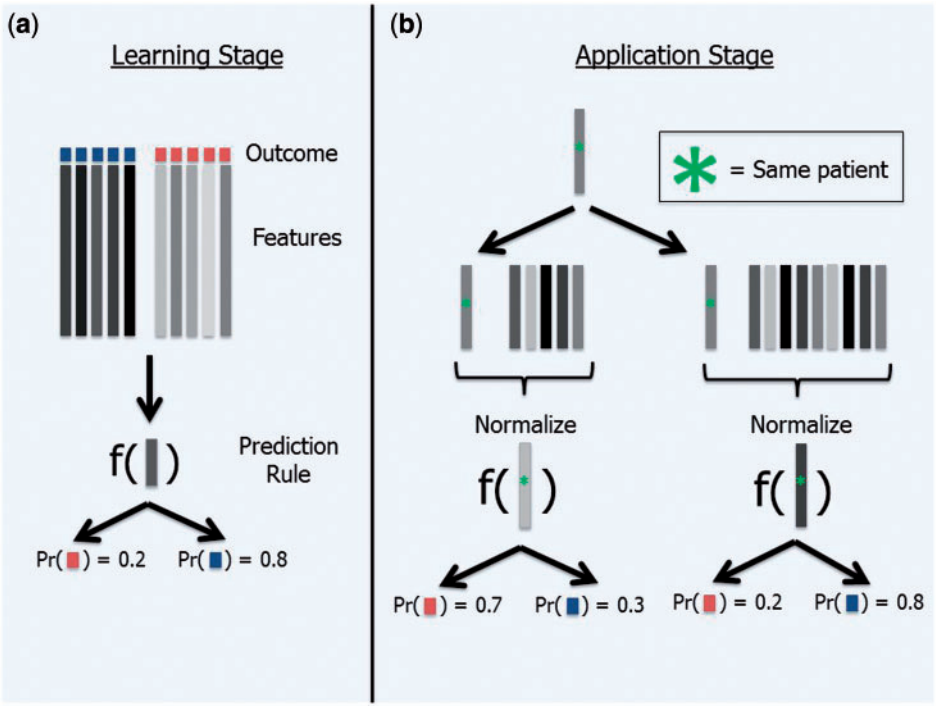

Fig. 1. A description of how test set bias can alter class prediction for an individual patient. In (**a**), we learn a model for predicting if a patient is in class R (red) or class B (blue). In our training data, the patients with darker gray features tend to be in class B, whereas the lighter gray patients are in class R. We develop a prediction rule from our training data and apply it to a new darker gray patient, and we see that he is likely to be classified to class B. In (**b**), we attempt to classify a single patient in the context of two different patient populations. We see that depending on the number and type of other patients in the population when we normalize the data, the resulting feature profile for our patient can be drastically different. This leads to different eventual classifications by our prediction rule. We contend that the ultimate classification of a patient should not depend on the characteristics of the test set but rather solely on the characteristics of the patient himself

The case we would like to explore is when the gene signature and prediction algorithm are 'locked down' and when there is no biological variation in the patient data. We are concerned with how much data transformation due to pre-processing and normalization affects classification. It is our assertion that steps taken to transform patient data for the purposes of *applying* a prediction algorithm should not alter the patient's eventual classification.

Some normalization methods (Bengtsson *et al.*, 2008; McCall *et al.*, 2010; Piccolo *et al.*, 2012) and some batch correction methods (Leek *et al.*, 2012; Parker *et al.*, 2014) have addressed this issue by normalizing each sample against a fixed, or 'frozen', set of representative samples. Unfortunately, these approaches can be applied only to specific platforms where large numbers of representative samples have been collected. This is especially relevant when custom chips are designed, as is the case in many clinical applications. There remain a large range of platforms for measuring gene expression in use by researchers (Barrett *et al.*, 2013), and single sample normalization methods are not currently available for many of these platforms. Additionally, methods such as quantile normalization and other forms of data scaling and transformation have become well known in the field and are often applied as standard steps in a data processing pipeline.

Even if single sample normalization methods were universally available, public measures of gene expression are frequently preprocessed using a range of methods for cleaning, normalization and analysis, resulting in a range of expression values for the same gene across different platforms (Allison *et al.*, 2006). A more tractable solution is to build gene signatures that do not rely on raw gene expression values. We propose using the ranks of genes instead of their raw expression values under the assumption that any transformation applied to the data is rank-preserving.

As a concrete example, we focus on the PAM50 signature for breast cancer subtyping (Parker *et al.*, 2009), which is used to assign patients with breast cancer to one of five molecular subtypes: Basal, Luminal A, Luminal B, Her2 and Normal. We show that when the number of patients in the test set changes, the predictions for a single patient may change dramatically. We also show that variation in patient populations being predicted upon leads to test set bias. Interestingly, PAM50 can be easily modified into a rank-based signature. We show that predictions from rank-based PAM50 are comparable to those from standard PAM50 and that predictions from rank-based PAM50 are invariant to test set bias.

Test set bias is a failure of reproducibility of a genomic signature. In other words, the same patient, with the same data and classification algorithm, may be assigned to different clinical groups. A similar failing resulted in the cancellation of clinical trials that used an irreproducible genomic signature to make chemotherapy decisions (The Cancer Letter, 2011). The implications of a patient's classification changing due to test set bias may be important clinically, financially and legally. In the example of PAM50, a patient's classification could affect a treatment or therapy decision. In other cases, an estimation of the patient's probability of survival may be too optimistic or pessimistic. The fundamental issue is that the patient's predicted quantity should be fully determined by the patient's genomic information, and the bias we will explore here is induced completely due to technical steps.

2 Materials and Methods

2.1 Study population and data

We collected and curated gene expression microarray data representing 28 independent studies (Haibe-Kains *et al.*, 2012). These datasets

Table 1. Baseline characteristics of curated dataset

Characteristic	Summary
N	6297
Age (years)	57.29 (13.42)
RFS (years)	7.22 (4.86)
Tumor size (cm)	2.52 (1.43)
Node	
+	1871
–	2857
NA	1569
Grade[a]	
1	525
2	1642
3	2226
NA	1904
ER	
+	3635
–	1556
NA	1106
PGR	
+	766
–	656
NA	4875
Her2	
+	496
–	1437
NA	4364
Subtype[b]	
Basal	1254
Her2	927
LumA	2007
LumB	1813
Normal	296

Her2, human epidermal growth factor receptor 2 status; node, whether or not cancer has spread to lymph nodes; PGR, progesterone receptor status; RFS, recurrence-free survival time. Age, RFS and tumor size are given as means with standard deviations.

[a]Because of the ambiguity of grade 2, we chose to build all prediction models for grades 1 and 3 only.

[b]Subtypes as predicted by PAM50.

spanned 15 different proprietary platform types and a variety of platform versions and included a range of commercial and private manufacturers, spanning Affymetrix, Illumina and Agilent as well as custom arrays. The data were collected from the Gene Expression Omnibus (Barrett *et al.*, 2013), ArrayExpress (Parkinson *et al.*, 2007), The University of North Carolina at Chapel Hill database (UNCDB), Stanford Microarray Database (SMD) and Journal and Authors' websites. Metadata were manually curated as previously described (Haibe-Kains *et al.*, 2012). Experiments ranged from 43 to 1992 patients, with a median of 131 patients and a total of 6297 patients (Table 1).

2.2 PAM model fitting

Prediction analysis of microarrays (PAM) (Tibshirani *et al.*, 2002) is a commonly used supervised learning approach for building prediction models using gene expression data from microarrays. We employed the pamr package (Hastie *et al.*, 2014) to fit a PAM model using R. Briefly, pamr takes class labels and microarray data and calculates an average gene expression profile, or centroid, for each class. It then shrinks the centroid to eliminate genes that do not contribute to explaining variability between classes. We then cross-validate to find an appropriate shrinkage threshold to maximize predictive accuracy of our model. We use this threshold to determine how many of the genes to keep in the predictor.

2.3 Normalization procedure

Normalization is accomplished through quantile rescaling as implemented in the genefu package (Haibe-Kains *et al.*, 2011). This scales each gene expression value x using specific quantiles from the expression data. First, a quantile q is chosen. Through examination of many microarray datasets, $q = 0.05$ was found to be robust. The expression values corresponding to the desired quantiles $q_1 = x_{\frac{q}{2}}$ and $q_2 = x_{1-\frac{q}{2}}$ are defined, and the scaled value $x' = \frac{x-q_1}{q_2-q_1}$ is calculated. In contrast to scaling by the maximum and minimum value, this approach is more robust to extreme outlying gene expression values.

This normalization procedure is applied internally when the intrinsic.cluster.predict function from the genefu package is used and the model's standardization ('std') parameter is set to 'robust'. For example, we can make PAM50 predictions using pre-packaged models in genefu called pam50 or pam50.robust. The gene centroid information is the same in both cases, but pam50 has std = 'none' and pam50.robust has std = 'robust'. This means that if we apply intrinsic.cluster.predict with pam50, the test data will not be normalized in any way, but if we use pam50.robust the quantile rescaling procedure described above will be applied.

2.4 Estimating test set bias

We used two approaches to estimate test set bias. When considering the PAM50 predictor, we simply applied the pre-defined prediction model from the genefu package (Haibe-Kains *et al.*, 2011) to make predictions on our data.

To train a PAM model, we used 10-fold cross-validation. We create a test set that is approximately 10% of the total data and use the remaining 90% to train the model. We use the internal cross-validation functions provided in the pamr package (Hastie *et al.*, 2014) to produce a shrinkage threshold and determine the number of genes necessary to make predictions. We then apply this predictor both in the test set, which comes from the same platform, and on other microarray datasets that used different platforms. This process is repeated within each of the cross-validation folds to get average prediction accuracies and standard deviations. When predicting tumor grade (1–3 with increasing severity), we restricted to patients graded 1 or 3 as grade 2 is considered to be ambiguous.

3 Results

3.1 Normalization makes patient predictions depend on other patients' data

Consider the PAM50 signature (Parker *et al.*, 2009). The class assignment for a new patient is made by calculating a measure of closeness between the new patient and the average patient profile in each possible class, then choosing the class that was closest to the sample. For example, PAM50 consists of 50 genes and predicts five classes, so each class centroid is a profile of the average expression of each of the 50 genes within that class. The authors used correlation as a measure of closeness for a given sample to each class centroid, i.e. correlation is calculated between the 50 genes in the patient sample and the 50 genes in each class centroid. This is the step that necessitates suitable rescaling of the test data before predictions are made.

We considered two scenarios, which illustrate how PAM50 can produce varying subtype predictions for a particular patient when

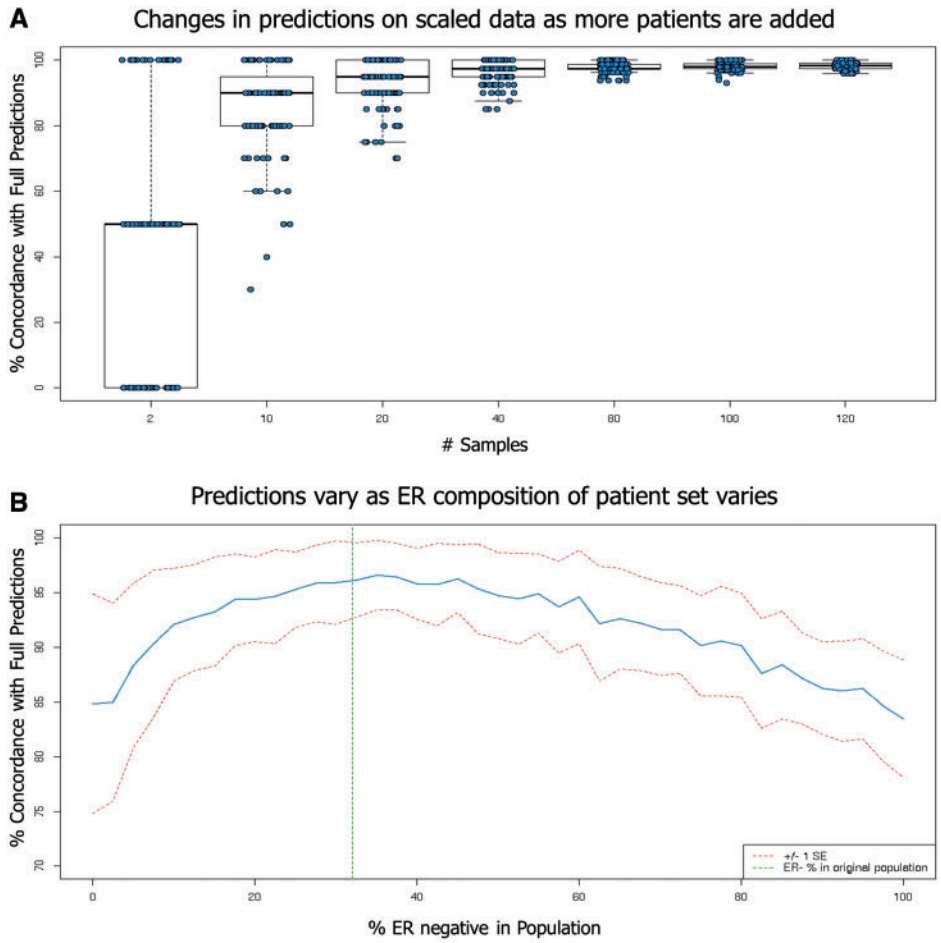

Fig. 2. Predictions for an individual patient can change depending on how many and what type of patients are included in the normalization step. (**A**) We first predicted the PAM50 subtype on an entire set of patients (Affymetrix hgu133plus2; GSE7390; $n = 198$). We then took 100 random samples of patient subsets ranging from 2 to 120 patients and predicted their subtypes with data normalization. We compared this newly predicted subtype to each patient's originally predicted subtype and calculated agreement. Actual data are jittered and overlayed on the boxplot. We find that there is significant variation in percent concordance when a small subset of patients is subtyped in comparison to the entire patient population. (**B**) From the same setup, we took 100 random samples each of 40 patients and varied the percentage of ER-positive and ER-negative patients in the sample. That is, 0% on the X-axis corresponds to 0% (0/40) ER-negative patients and 100% (40/40) ER-positive patients in the sample. We then predicted subtypes on this subset and compared these newly predicted subtypes to the original predictions. The average concordance is plotted with ± 1SE bands. We note that the original population is 32% ER negative (dashed green line), which is where we see close to maximal concordance

the data for other patients used in normalization varies. We used data from GSE7390 ($n = 198$), an experiment conducted using the Affymetrix hgu133plus2 microarray. In each experiment, we normalized the gene expression measurements in the test set to fall between 0 and 1.

First, we created predictions where we normalized all patients together. Then we calculated predictions for the same patients when normalized in smaller groups ($n = 2,10,20,40,80,100,120$) and measured the agreement between the predictions for the exact same patient when normalized with all patients versus a smaller subset of patients. When normalized in small batches, the predictions for the same patient changed compared with the case where all patients were normalized together (Fig. 2A).

Next we predicted on patient populations that varied in the distribution of estrogen receptor (ER) status, which is an important factor in breast cancer prognosis and treatment. Again we applied the PAM50 predictor to the entire test set. Then we created subsets of the entire test set with differing percentages of ER-negative patients and applied the predictor to each subset. When the percentage of ER-negative patients in the subset matched the percentage in the entire test set, patient subtypes best agreed with the original predictions on the entire test set. However, when the ER status of the other patients in the test set varied, the predictions for the same patient were often different (Fig. 2B).

3.2 Using gene ranks with unnormalized data produces comparable accuracy

When PAM50 was proposed, the authors chose to calculate similarity based on Spearman correlation (Parker *et al.*, 2009). Spearman correlation finds the correlation between the *ranks* of the two sets of gene expression measurements rather than correlation between the actual values. We hypothesized that this rank-based prediction would be immune to some changes of scale across platforms and other platform-specific artifacts. With traditional signatures, these are precisely the reasons why normalization is necessary. To examine this preliminarily, we re-ran the process from the previous section but simply did not normalize the data and relied on the internal rank-based correlation calculation. We recreated Figure 2A and B when the data were 'unscaled'. These appear as Supplementary

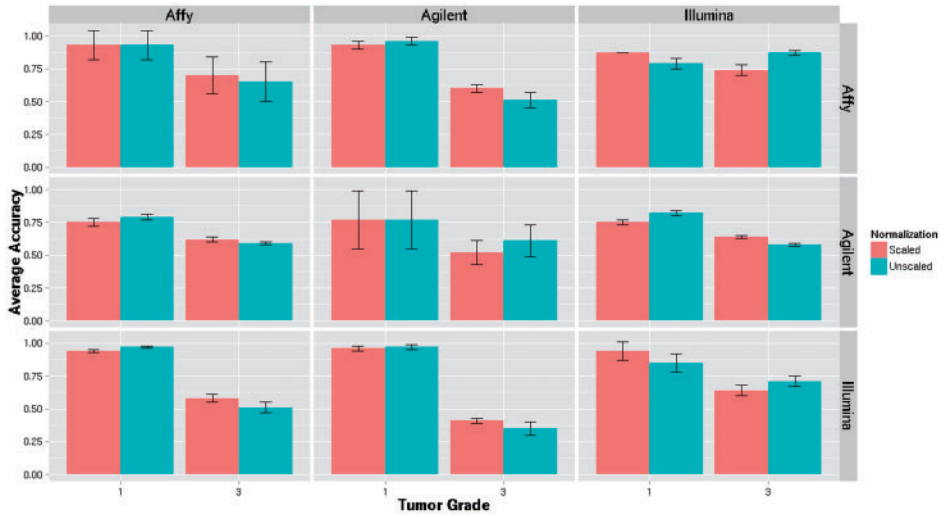

Fig. 3. Average accuracy of scaled and unscaled predictions over different training and testing sets we trained a PAM model to predict tumor grade (either grade 1 or 3) using 10-fold cross-validation on one Affymetrix (GSE7390), Agilent (ISDB10845) and Illumina (ISDB10278) dataset each. The rows represent upon which platform each model was trained, and the columns represent upon which platform each trained model was applied to make predictions. To get average accuracy and standard deviations (error bars) for a particular platform, we use the model generated under each fold of the cross-validation to make predictions on the remaining test set of the same platform as well as the two other platforms. We applied this model after normalizing ('scaled') the data and after leaving it unnormalized ('unscaled'). We found that the accuracies for predicting grade were similar whether the data were normalized or unnormalized

Figures SI and SII, and they show that the predictions remain constant as sample size and ER status vary when the data are unnormalized and a rank-based metric is employed.

To further evaluate this hypothesis, we used the previously proposed PAM signature-building procedure (Tibshirani *et al.*, 2002) to build a genomic signature to predict tumor grade (a clinical quantity indicating severity) using three datasets measured on different platforms: GSE7390 (Affymetrix; $n = 198$), ISDB10845 (Agilent; $n = 337$) and ISDB10278 (Illumina; $n = 1992$). We used 10-fold cross-validation to train a model on a particular dataset, made predictions on the testing portion of that dataset and applied the trained model to the two remaining datasets, which represent two different platforms. We averaged over the 10-folds in each case to obtain mean accuracy and standard deviation.

To make predictions, we used Spearman correlation to mimic how the PAM50 signature is used (Parker *et al.*, 2009). We predicted new patient samples using our PAM signature for grade both with and without normalization. The same set of genes and prediction algorithm are used in both cases—the only difference is that in the former we normalize the test set patient data, and in the latter, we leave it unnormalized. We observed that the normalized and unnormalized predictors performed similarly across platforms (Fig. 3).

Within-platform (Affy-Affy, Agilent-Agilent, Illumina-Illumina in Fig. 3), there is no appreciable difference in the average accuracy of predictions when the test data are normalized or unnormalized. For Affy, the grade 1 and 3 average accuracies and standard deviations (represented by error bars in the figure) when the data are normalized are 0.92 (0.13) and 0.67 (0.17), respectively, when compared with 0.92 (0.13) and 0.65 (0.16) when the data are unnormalized. For Agilent, the relevant figures are 0.72 (0.32); 0.56 (0.05) for normalized versus 0.72 (0.32); 0.65 (0.09) for unnormalized and for Illumina 0.92 (0.06); 0.65 (0.05) versus 0.84 (0.08); 0.71 (0.06). In all cases, the ranges of the unnormalized average accuracies substantially overlap those of the normalized average accuracies. Results across platforms (the off-diagonal grid entries in the figure) tell a similar story. It is the case that if the scaled predictor performs better on grade 1 than the unscaled, then the opposite will be true for grade 3 (see e.g. the Agilent-Illumina result). This is due to the fact that patients can be classified as either grade 1 or 3, so if the unscaled version predicts more grade 3 than grade 1, the change in the respective accuracies will be proportional. This analysis suggests that using the PAM predictor for grade with Spearman correlation and without normalizing the test set data produces similar predictive accuracy to when the test set data are normalized.

4 Discussion

We found that breast cancer tumor subtype predictions varied for the same patient when the data for that patient were processed using differing numbers of patient sets and patient sets had varying distributions of key characteristics (ER status). This is undesirable behavior for a prediction algorithm, as the same patient should always be assigned the same prediction assuming their genomic data do not change. The fact that sample size affects normalized data values is unsurprising, but the fact that classifications varied by how many patients were ER− in the test set speaks to the generalizability of an algorithm. Ideally, the test set should be 'similar' in composition to the dataset upon which a classification algorithm was trained. The result in Figure 2B is undoubtedly related to the fact that ER+ patients are different in terms of gene expression from ER− patients, but we see that even slight perturbations in the ER composition of the subpopulation can affect patient classifications. This raises the question of how similar the test set needs to be to the training data for classifications to be trusted when the test data are normalized.

The PAM50 signature uses Spearman correlation to assess distances when making predictions, so we leveraged this by comparing how a PAM signature using Spearman correlation predicts tumor grade outcomes with and without normalization. We found the results to be comparable, but the unnormalized approach guarantees the same prediction for the same patient every time. A gene signature that employs rank-based features or makes other rank-based calculations is one robust approach to avoiding test set bias. Although all gene signature classifiers do not necessarily have a

completely rank-based mode as PAM50 does, the broader implication of this result is that one may try to build predictors that operate only on the ranks of data, thereby bypassing the need for any normalization step when predicting on a test set.

Acknowledgements

This study used data generated by METABRIC; we thank the British Columbia Cancer Agency Branch for sharing these invaluable data with the scientific community. We acknowledge the great support from the InSilicoDB team for storage and programmatic access to our compendium of breast cancer microarray datasets.

Funding

B. Haibe-Kains was supported by a Cancer Research Society Operating Grant (Canada). JTL and PP were partially supported by NIH Grant U54CA151838.

Conflict of Interest: none declared.

References

Allison,D.B. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, **7**, 55–65.

Baggerly,K.A. *et al*. (2005) Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *J. Natl. Cancer Inst.*, **97**, 307–309.

Barrett,T. *et al*. (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**(Database issue), D991–D995.

Bengtsson,H. *et al*. (2008) aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. *Technical report 745*. Department of Statistics, University of California, Berkeley.

Glas,A.M. *et al*. (2006) Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*, **7**, 278.

Haibe-Kains,B. *et al*. (2014) *genefu: Relevant Functions for Gene Expression Analysis, Especially in Breast Cancer*. R package version 1.16.0, http://www.pmgenomics.ca/bhklab/.

Haibe-Kains,B. *et al*. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *J. Natl. Cancer Inst.*, **104**, 311–325.

Hastie,T. *et al*. (2014) pamr: Pam: Prediction Analysis for Microarrays. R package version 1.55. http://CRAN.R-project.org/package=pamr.

Leek,J.T. *et al*. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Letter,T.C. (2011) *Duke Accepts Potti Resignation; Retraction Process Initiated with Nature Medicine*.

Lusa,L. *et al*. (2007) Challenges in projecting clustering results across gene expression profiling datasets. *J. Natl Cancer Inst.*, **99**, 1715–1723.

Majewski,I.J. and Bernards,R. (2011) Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nat. Med.*, **17**, 304–312.

McCall,M.N. *et al*. (2010) Frozen robust multiarray analysis (frma). *Biostatistics*, **11**, 242–253.

Michiels,S. *et al*. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

Parker,H.S. *et al*. (2014) Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ*, **2**, e561. DOI: 10.7717/peerj.561.

Parker,J.S. *et al*. (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Parkinson,H. *et al*. (2007) ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**(Database issue), D747–D750.

Petricoin,E.F. *et al*. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Piccolo,S.R. *et al*. (2012) A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, **100**, 337–344.

Sebastiani,P. *et al*. (2010) Genetic signatures of exceptional longevity in humans. *Science*, **2010** [Epub ahead of print, doi: 10.1126/science, July 1, 2010].

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.