Bioinformatics, 31(6), 2015, 912–918
doi: 10.1093/bioinformatics/btu736
Advance Access Publication Date: 6 November 2014
Original Paper

Data and text mining

Development of a robust classifier for quality control of reverse-phase protein arrays

Zhenlin Ju[1], Wenbin Liu[1], Paul L. Roebuck[1], Doris R. Siwak[2], Nianxiang Zhang[1], Yiling Lu[2], Michael A. Davies[2,3], Rehan Akbani[1], John N. Weinstein[1,2], Gordon B. Mills[2] and Kevin R. Coombes[1,*,†]

[1]Department of Bioinformatics and Computational Biology, [2]Department of Systems Biology and [3]Department of Melanoma Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

*To whom correspondence should be addressed.
†Present address: Departments of Biomedical Informatics, The Ohio State University College of Medicine, 340E Lincoln Tower, 1800 Cannon Drive, Columbus, OH 43210, USA

Associate Editor: Jonathan Wren

Received on May 1, 2014; revised on October 7, 2014; accepted on November 3, 2014

Abstract

Motivation: High-throughput reverse-phase protein array (RPPA) technology allows for the parallel measurement of protein expression levels in approximately 1000 samples. However, the many steps required in the complex protocol (sample lysate preparation, slide printing, hybridization, washing and amplified detection) may create substantial variability in data quality. We are not aware of any other quality control algorithm that is tuned to the special characteristics of RPPAs.
Results: We have developed a novel classifier for quality control of RPPA experiments using a generalized linear model and logistic function. The outcome of the classifier, ranging from 0 to 1, is defined as the probability that a slide is of good quality. After training, we tested the classifier using two independent validation datasets. We conclude that the classifier can distinguish RPPA slides of good quality from those of poor quality sufficiently well such that normalization schemes, protein expression patterns and advanced biological analyses will not be drastically impacted by erroneous measurements or systematic variations.
Availability and implementation: The classifier, implemented in the "SuperCurve" R package, can be freely downloaded at http://bioinformatics.mdanderson.org/main/OOMPA:Overview or http://r-forge.r-project.org/projects/supercurve/. The data used to develop and validate the classifier are available at http://bioinformatics.mdanderson.org/MOAR.
Contact: Kevin.Coombes@osumc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The majority of known biological effector molecules, diagnostic markers and pharmaceutical targets are proteins (Nishizuka *et al.*, 2003). Historically, investigations have focused on single proteins. The ability to profile the expression of multiple proteins has added a major new dimension of information on biological and clinical processes. A technology that accomplishes that is reverse-phase protein array (RPPA). RPPA offers a powerful tool for defining more than 100 protein expression profiles for over a 1000 samples simultaneously in an experiment. Typically, an RPPA experiment is performed by extracting protein lysates from biological samples, making a series of dilutions of the lysates, printing the lysates onto nitrocellulose-coated glass slides, hybridizing the samples with primary and enzyme-tagged secondary antibodies, then adding a substrate of the enzyme to generate a colorimetric signal. The signal intensities, which reflect the abundance of the corresponding proteins in the hybridized samples, are then detected and quantified by any number of approaches, such as those used by the software MicroVigene (VigeneTech, Inc., http://www.vigenetech.com/MicroVigene.htm). This software computes the mean net and mean total values of the signals. To ensure data validity, a positive control derived from the same lysate extract of the mixed cell lines is repeatedly spotted on each slide. Each sample and positive control on an RPPA slide usually has five dilution steps (undiluted, 2×, 4×, 8× and 16×). A typical RPPA slide contains 48 subgrids, and each subgrid has the dilution series of 22 samples and 2 positive controls. Therefore, an RPPA slide contains up to 5760 spots, which can accommodate up to 1152 samples and controls (Fig. 1). A large amount of data can be obtained from a single experiment. Thus, an experiment requires extensive data handling, including quality control (QC).

RPPA technology has been used in a variety of studies, such as those involving cellular signaling pathway activation (Davies *et al.*, 2009; Paweletz *et al.*, 2001), discovery of biomarkers of treatment response and survival (Carey *et al.*, 2010; Grote *et al.*, 2008; Liang *et al.*, 2012; Park *et al.*, 2010; Tsao *et al.*, 2010), disease classification for diagnostic and prognostic purposes (Tibes *et al.*, 2006), and identification of new drug targets (Chen *et al.*, 2011; Nishizuka *et al.*, 2003; Tsao *et al.*, 2010). Most studies that have used RPPA have concentrated on biological analyses; however, to our knowledge, no study has been published that explores QC for RPPA slides.

It is common for an RPPA experiment to result in some slides of poor quality due to experimental variation that arises even under optimal conditions. To avoid poor RPPA images and subsequent spurious data—which may have adverse impact on normalization, lead to wrong conclusions, take extra time to process and require extra space to store—it is crucial to exclude slides with poor quality in the early stages of data processing. Current QC strategies are based on a visual examination of slide images or correlation coefficient analysis on a limited number of replicates if there are replicates on a slide. However, when analysing a large number of slides, visual examination is labor-intensive and time-consuming. Moreover, the QC criteria for visual examination may vary from examiner to examiner, which may lead to inconsistent or inaccurate results. The alternative, correlation coefficient analysis, indicates a predictive relationship between two variables, but not reproducibility or reliability.

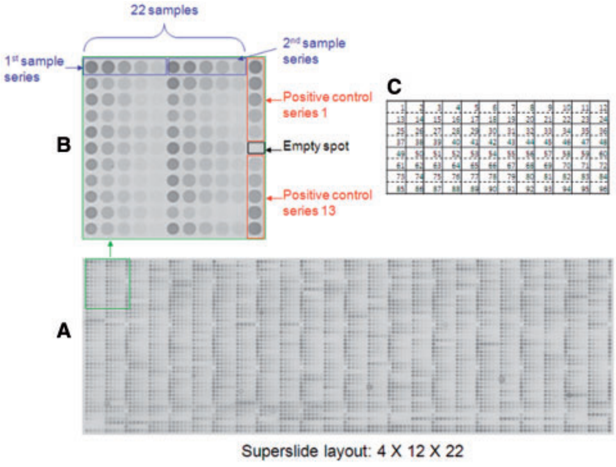
22 samples
1st sample series
2nd sample series
Positive control series 1
Empty spot
Positive control series 13
B
C
A
Superslide layout: 4 X 12 X 22

Fig. 1. Example layout of a superslide. (**A**) The entire superslide, containing 48 (4 × 12) subgrids, with each grid having 22 sample series, 2 positive control series and 1 empty spot. Each series has five spots representing five dilution steps. Each subgrid has 121 spots. (**B**) A subgrid, showing the locations of the sample series, positive control series and empty spot. (**C**) Overall view of 96 replicated positive control series on a superslide

In general, the development of a QC model consists of defining descriptive categories, selecting features and constructing the model. The descriptive categories are quality criteria, with predictors being used to measure the criteria. In this study, each qualitative category corresponds to the nature of an RPPA slide. In practice, biologists are interested in excluding unreliable slides from further analysis by dividing the slides into two categories: one category that consists of slides that are excluded from further analysis (referred to as slides of poor quality hereafter) and a second category that consists of slides that are used in the actual data analysis (referred to as slides of good and fair quality hereafter). The classification problem arises when the categories established for the slides vary from biologist to biologist. To tackle that problem, we define the quality of an RPPA slide using an exhaustive set of quantitative predictors in a generalized linear model (GLM; Dobson and Barnett, 2008; Hardin and Hilbe, 2007) and a logistic function by which an outcome of a GLM calculation is transformed to the probability that the slide is of good quality. We demonstrate here the quantitative features, model training and model validation of our RPPA QC model.

2 Datasets

Three independent datasets, containing 138, 117 and 174 slides, respectively, were used in this study. One set ($n = 138$) was used as the training set, and the other two sets were used as the validation sets. Based on visual examinations of the images, the slides of the three sets were assigned to three quality categories, good, fair and poor, by two or three experts who have conducted RPPA experiments and data analyses for years. Slides categorized as ‘poor’ were of unacceptable quality, and those categorized as ‘good’ and ‘fair’ were of acceptable quality. As the reliability of an RPPA experiment is affected by many factors, it is not always straightforward to perform a visual assessment of slide quality. Therefore, using their prior knowledge of RPPA experiments and slide quality, the experts evaluated slide quality based on four factors: the shape and size of a spot, the intensity of a spot, the background intensity and uneven patches on the slide and variations in the positive controls.

Spot shape and size: Spots are expected to be roughly round in shape and equal in size, but irregular morphologies such as bleeding, starring, etc., and variations in size may occur during RPPA slide printing. The causes of irregular shape include scratches on the slide surface, uneven coating of nitrocellulose on the slide and uneven distribution of sample lysate in the spot. The deviation from a standard size can be caused by impurities and debris in the sample lysate and printing solution; unlevel slides, such that the printing needles do not make adequate contact with the array surface; or samples evaporating during printing. Too large a spot can lead to problems with detecting and quantifying the surrounding spots.

Spot intensity: Signal intensity is a universally accepted criterion for the evaluation of array quality. A weak signal makes it difficult to distinguish between the actual signal and background noise. In contrast, saturated signals do not measure biological variation in levels. In addition, both weak and saturated signals pose challenges to the sensitivity limits of an array scanner.

Background intensity and uneven patches: An ideal RPPA image is one in which sample spots are visualized without background or with an evenly distributed light background. An uneven, dark background makes it hard to detect and quantify the true spot intensity. Background patches with various levels of darkness may cause a biased estimation of the true spot intensity. The sources of the background signal can be background staining from the nitrocellulose coating of the glass slide and contamination from the hybridization and washing procedures.

Varia tion in the positive controls: The intensities of positive control spots are expected to be equal because they are technical replicates. A visualized intensity variation among the positive control spots indicates uneven hybridization, an uneven background or irreproducible technical replicates.

We used Cohen's kappa test (Cohen, 1968) to evaluate the agreement in the slide quality categorization from the different experts for each dataset. The experts significantly agreed on their visual categorizations, with an average *P*-value of 4.22E−07 for Cohen's kappa (*P* values ranged from 2.20E−16 to 2.86E−06). To finalize the categorization of the slides, we assigned a score to each category (3 was good, 2 was fair and 1 was poor), averaged the scores from the different experts for each slide, and used 1.5 and 2.5 as cutoffs. A slide with an average score smaller than or equal to 1.5 was classified as a poor slide; a slide with an average score between 1.5 and 2.5 was a fair slide and a slide with an average score larger than or equal to 2.5 was a good slide. As a result, the training set had 50 good, 64 fair and 24 poor slides. One of the validation sets ($n = 117$) had 57 good, 40 fair and 20 poor slides. The other validation set ($n = 174$) contained 29 good, 75 fair and 70 poor slides (Supplementary Table S1). The process that the experts used to set up the criteria and perform the quality evaluation was completed prior to and independent from model development.

3 Quantitative features

There are many features that reflect the quality of a slide. In this section, we describe how we used the mean net and/or mean total to define these features and explain why we selected the features.

3.1 Dynamic range of signal intensity of a sample

Dynamic range (DR) is the measure of the difference between the smallest and largest possible mean net intensities in each dilution step for the samples. We calculated DR as follows:

$$\mathrm{DR}_{ni} = \max\,(I_{ni}) - \min\,(I_{ni}), \tag{1}$$

where DR_{ni} is the difference between the largest and smallest intensities of the *i*th dilution step; *i* is from 1 to 5, representing each of the dilution steps (the undiluted, 2×, 4×, 8× or 16× dilution step); *n* is the number of slides included in the RPPA set and I_{ni} is a vector containing the 'mean net' (see section 3.2 for definition) intensities of the *i*th dilution step.

$$I_{ni} = \{m_{ni1}, m_{ni2}, m_{ni3},\ldots\ldots, m_{nik}\}, \tag{2}$$

where m_{nik} is the signal intensity of the *k*th sample of the *i*th dilution step in the *n*th slide.

The samples spotted on a slide often consist of a wide spectrum of tissue types, tumor types, tumor stages or grades and/or treatments, etc. The variation in protein expression should be fairly large among the groups of samples. Therefore, large DRs are expected. A small DR can be due to weak or saturated signals. If the signals on a slide are overall weak to non-existent or strong to saturated, the true protein expression of the samples may not be reflected. In addition, weak signals make it difficult to distinguish foreground signals from background noise, while saturated signals do not reflect the true amounts of protein concentration. Experimental factors that may cause weak or saturated signals include extremely high or low amounts of protein in the spot; suboptimal concentration of the antibody; improper dilution of the sample lysate; incomplete or over-hybridization; incomplete or over-washing and suboptimal sensitivity of the scanner. Non-specific binding may contribute to the signal saturation as well. Regardless of the actual cause, weak or saturated signal intensities do not reflect biological variation and slides with weak or saturated signal intensities are considered less reliable and of poor quality.

3.2 Background–to-signal ratio

The signal intensity collected by MicroVigene software (http://www.vigenetech.com/MicroVigene.htm) includes the values of the mean total and mean net for each spot. The mean total is the mean value of the spot intensity before background subtraction. The mean net is the spot mean minus the background intensity. The background intensity is determined by the percentile of spot background intensities. Using the mean total and mean net, we computed the background–to-signal ratio (BSR) for each sample spot, as follows:

$$\mathrm{BSR}_{nik} = \frac{M_{nik} - m_{nik}}{m_{nik}}, \tag{3}$$

where M_{nik} and m_{nik} represent the signal intensities of the mean total and mean net, respectively, of the *k*th sample of the *i*th dilution step in the *n*th slide, and *i* is from 1 to 5, representing the five dilution steps. Therefore, the BSR measures the fraction of the background in the total signal, indicating how much a signal has been affected by background noise and debris. For each slide, we used the BSR to compute an index that was the percentage of spots with BSRs smaller than 1:

$$Pg_j = \frac{n_j}{N_j}, \tag{4}$$

where Pg_j is the index, n_j is the number of spots with BSRs smaller than 1 and N_j is the total number of spots on slide *j*. This index is a reflection of the percentage of good sample spots with foreground signals that are higher than the background noise. The larger the index, the larger the percentage of good sample spots and the more reliable the slide. Therefore, the index is a useful feature for assessing slide quality.

3.3 Slopes of each dilution step and each dilution series of the positive control

We used a single positive control based on a pool of cell lines to provide a wide number of positive control spots. The single positive control was spotted on the slide as 96 replicated dilution series, with each series having five dilution steps: undiluted, 2×, 4×, 8× and16× dilution; thus, each dilution step consisted of 96 replicated spots (Fig. 1). If perfectly reproduced, the signal intensities of the 96 replicated spots in each dilution step should be identical and form a straight horizontal line with a slope of zero when plotted in a 2D plot (Supplementary Figure S1A). However, the reality was that the signal intensities of every dilution step had certain degrees of variation among the replicates. When fitted with a linear model,

$$m_i = a + bx, \tag{5}$$

m_i is a set containing the mean net intensities of the positive control replicates in the *i*th dilution step, *i* is from 1 to 5, *x* is a vector containing the sequential numbers 1 to 96, a is the intercept and b is the slope of the dilution step (SStep). A non-zero b reflects the deviation of the observed line from a straight horizontal line, which indicates that the spots are not reproducible. Therefore, the SStep is related to slide quality.

In addition, if the dilution is exact, the protein concentration will be reduced by half sequentially in a series of five dilution steps (SSeries), and the signal intensities will be reduced by half sequentially in the SSeries. When fitted with a linear model and plotted in a 2D plot, the log-transformed intensities of an exact dilution series will generate a linear line with a standard slope of -0.6931 (Supplementary Figure S1B). However, in this analysis, most of the experimental lines were not aligned with the standard line. The difference between the values of an observed slope and the standard slope may be caused by inaccurate lysate dilutions, uneven hybridization and/or uneven background. The larger the difference, the lower is the reliability of the slide.

3.4 Coefficient of variation of each dilution step of the positive control

To monitor the intensity variation among the 96 replicate spots, we used a coefficient of variation (CV), which we computed as

$$\mathrm{CV}_i = \frac{\sigma_i}{\mu_i}, \tag{6}$$

where σ_i is the standard deviation and μ_i is the mean of the mean net values of the signal intensities of the positive control in the ith dilution step. Here, i is from 1 to 5, representing each of the five dilution steps, and μ_i is the mean of the 96 replicated mean net values.

The CV is optimal for determining a normalized measure of the dispersion of a data distribution (Hendricks and Robey, 1936). Because of the differences in the nature of protein expression levels, the mean of the signal intensities may vary widely from slide to slide (or protein to protein); thus, the CV is appropriate for comparing the variation among slides. The CV would be zero if the replicates were perfectly reproduced. The larger the CV, the greater the dispersion of the data points, which indicates poor reproducibility.

4 Generalized linear model

The GLM was chosen to fit the predictors. The dependent variable in the GLM is the score assigned to the slide quality (3 for good quality in a slide, 2 for fair or 1 for poor). We assume that the error distribution of the GLM follows a Gaussian distribution (Dobson and Barnett, 2008; Hardin and Hilbe, 2007). The greatest advantage of the GLM is that it generalizes linear regression by correlating a linear model to the dependent variable via a link function (‘identity’ in this study) and by correlating the magnitude of the variance of each independent variable to be a function of its predicted value.

We used the fitted GLM as a logit to a logistic function, by which the numeric outcomes generated by the GLM were transformed into probabilities. The classifier is then defined by the logistic function,

$$\theta = \frac{e^{(z)}}{1 + e^{(z)}}, \tag{7}$$

where θ represents the probability of the occurrence of an event, e is the base of the natural logarithm ($\sim$2.718) and z is the logit that measures the total contribution of all the predictor variables used in the model. Here, z is the GLM, denoted as

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_k x_k, \tag{8}$$

and specifies the linear relationship between z and a set of k predictor variables (x). β_0 is the intercept and $\beta_{1,}\ \beta_{2,}\ldots\beta_k$ are the regression coefficients of the predictors.

5 Data analysis

All computations and modeling were performed using the open-source software R, version 2.12.1 (http://www.r-project.org/). For the GLM, the R function ‘glm’ was used for model fitting and ‘stepAIC’ for stepwise model selection. The GLM was then used as a logit to the logistic function (7), computing the probability of good slide quality.

6 Results

6.1 Predictors and logit

Using the above equations, we computed five DRs (using 1 and 2), one index (percentage of good sample spots, using 3 and 4), five SSteps (5), one SSeries (5) and five CVs (6) for each slide of the training set. Therefore, a total of 17 features were available for predictor selection. The slides of poor quality had obviously smaller DRs (Supplementary Figure S2) and smaller indexes, which reflected the smaller percentages of good sample spots (Supplementary Figure S3), but relatively larger CVs (Supplementary Figure S4) and SSeries (Supplementary Figure S5) of the poor quality slides compared with those features in the slides of good or fair quality. The SSteps varied (Supplementary Figure S6): compared with those of the good or fair slides, the median slopes of the poor slides were larger in the undiluted positive control spots, but smaller in the diluted positive control spots. Moreover, the slopes of the poor slides were more widely distributed than those of the good or fair slides.

In order to build a logit, we created a quantitative response variable by assigning a score to each category: 3 to good, 2 to fair and 1 to poor, and fit a GLM to the 17 features to measure their contributions to the response variable. Because the utility of features in slide quality assessment is largely unknown, a natural approach is to try to use the most informative features in the GLM. Using all the features as predictors of slide quality may be overly conservative. Therefore, we utilized Akaike’s information criterion (AIC; Akaike, 1974) to estimate the goodness of fit of the GLM, and used AIC-stepwise selection to sequentially remove the less informative features from the model until a model with the lowest AIC was obtained (AIC = 201.5; Supplementary Table 2 lists the features that were removed). After exhaustive selection, eight predictors remained in the GLM: the SSeries; the CVs of the undiluted, 2 × diluted and 8 × diluted positive control spots; the SStep of the 16 × diluted positive control spot; the DRs of the 2 × and 8 × diluted samples and the index of the percentage of good sample spots. The contribution of these predictors to the response variable of the GLM is summarized in Table 1.

6.2 Classifier

The classifier was developed by using the GLM (8) as a logit in the logistic function (7). The outcome of the classifier is the probability (θ) of good slide quality. To estimate the performance of the classifier, we grouped the three categories of slide quality in the training set into two categories: combining good and fair as a single category

Table 1. Contribution of predictors to GLM

Predictor	Estimate	Standard error	Student t-test	Probability Pr ($>\|t\|$)
β_0[a]	3.013E+00	3.14E−01	9.60	2.00E−16
SSeries[b]	−9.585E−01	6.11E−01	−1.57	1.19E−01
CV0[c]	−2.151E+01	9.01E−00	−2.38	1.85E−02
CV2[d]	−4.306E+01	1.49E+01	−2.88	4.64E−03
CV8[e]	−1.929E+01	5.49E−00	−3.52	6.08E−04
SStep16[f]	−1.574E−02	6.65E−03	−2.37	1.95E−02
DR2[g]	3.885E−05	2.01E−05	1.93	5.61E−02
DR8[h]	−4.131E−05	1.79E−05	−2.31	2.22E−02
Pg[i]	1.271E−02	5.38E−03	2.36	1.96E−02

[a] Intercept in the GLM.
[b] Slope of the positive control dilution series.
[c] Undiluted positive control.
[d] 2× diluted positive control.
[e] 8× diluted positive control.
[f] Slope of the 16× diluted positive control.
[g] DR of the 2× diluted samples.
[h] DR of the 8× diluted samples.
[i] Percentage of good sample spots.

of good and retaining poor as the second category. We then performed receiver operating characteristic (ROC) curve analysis (Fig. 2). The ROC for the Bernoulli classifier showed that the area under the curve (AUC) was 0.96 (Fig. 2A), indicating that the classifier's performance was excellent. As θ is a continuous value between zero (poor) and one (perfect quality), it is sometimes difficult to determine an appropriate cutoff below which the slides are of poor quality. We therefore plotted the predicted probabilities and observed that all the good and most of the fair slides had probabilities of larger than or equal to 0.8 (Fig. 2B). Using 0.8 as the cutoff, we found that 12 slides were classified as poor and 126 as good. In order to test the accuracy of the classification, we performed cross-table analysis between the predicted and observed groups of the slides and found that the accuracy was 90% for the training set.

6.3 Validation of the classifier

We used the model to compute the probability for two independent validation datasets ($n = 174$ and 117, respectively; Table 2). The ROC curves showed that the classifier performed well, with an AUC of 0.95 and 0.93 in validation sets 1 and 2, respectively (Fig. 2C and E). The cutoff of 0.8 defined by the training set sufficiently categorized the slides of the validation sets into the good or poor group, with prediction accuracies of 84 and 87%, respectively.

As there is no gold standard for visual evaluation, to minimize the degree of disagreement between examiners that was caused by human error, the examiners sat side by side and closely examined the images and signal intensity files of the slides when the quality estimations were inconsistent between the classifier and the visual evaluation. This primarily concerned slides that the examiners had defined as having poor quality. The examiners then concluded that visual evaluation is less sensitive than the classifier because some light spots seemed 'invisible' to the human eye, but were detected by a slide scanner (Fig. 3A), and background patches on a slide made spots seem saturated to the human eye and thus prevented the examiners from making an accurate judgment (Fig. 3B). The classifier recovered some of the slides that had been defined as being of poor quality by the examiners. The slides that were identified as having good quality by the examiners but categorized as having poor quality by the classifier mostly had observable signals on a relatively clear background, but had either weak positive control spots or saturated sample signals that were not responsive to the sample lysate dilution (Fig. 3C). After discussion, the examiners agreed with the classifier on most of the discrepancies. As a result, the predictive accuracy of the classifier increased dramatically (up to 97%).

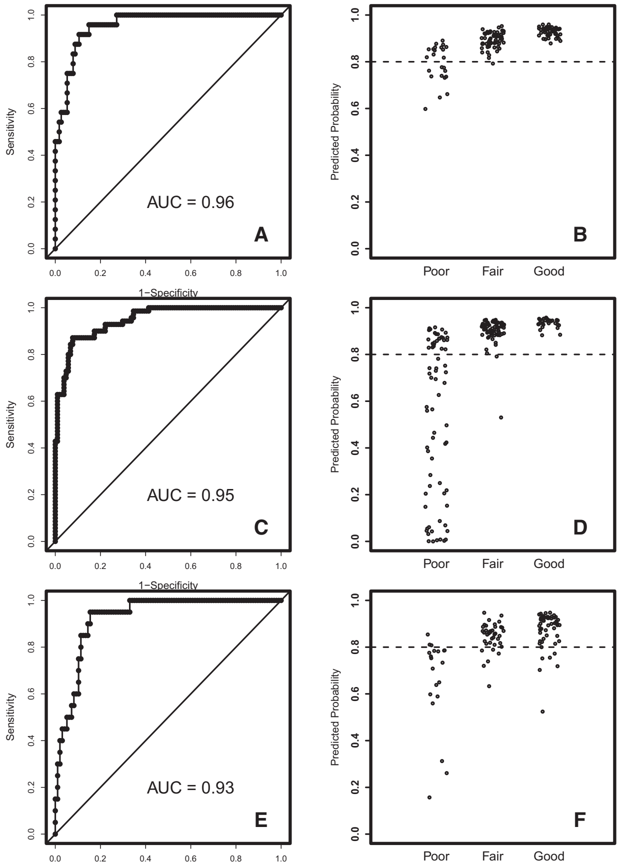

Fig. 2. QC model performance. ROC curves and distribution of slide quality probability of the training set (**A, B**); validation set 1 (**C, D**); validation set 2 (**E, F**). *X*-axes of graphs *B*, *D* and *E* represent slide quality categories as determined by visual assessment of the examiners; *Y*-axes are model-predicted probability of slide quality. The larger the probability, the better is the slide quality. Based on 0.8 probability cutoff derived from the training set, the model-predicted slide quality excellently in the two validation datasets, with an AUC of 0.95 and 0.93 for validation sets 1 and 2, respectively (C and E)

Table 2. Model performance

Dataset	No. of samples	No. of slides	Model performance			
			AUC[a]	PPV[b]	NPV[c]	Accuracy[d] (%)
Training set	1056	138	0.96	0.90	0.92	90
Validation set 1	1056	174	0.95	0.80	0.96	84
				0.97[e]	0.98[e]	97[e]
Validation set 2	1056	117	0.93	0.96	0.59	87

[a] AUC derived from ROC analysis.
[b] Positive predictive value.
[c] Negative predictive value.
[d] Prediction accuracy (%).
[e] Model performance after the examiners sat side by side and re-evaluated the slides.

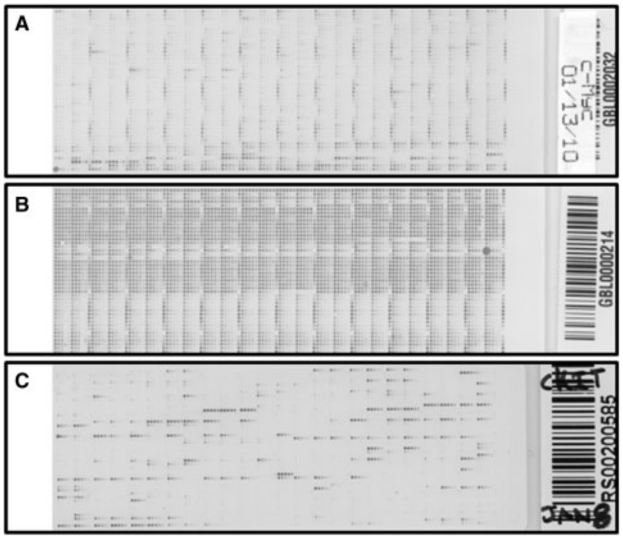

Fig. 3. Example RPPA slides. (**A, B**) Categorized as poor quality by the examiners, but as good quality by the classifier. The light spots seemed 'invisible' and the background patches made other spots seem saturated to the human eye, which prevented the examiners from classifying slides *A* and *B* as good slides. Spot-finding software is able to quantify these spots and provide acceptable data for further analysis; thus, the QC model defined them as good slides. (**C**) Categorized as good quality by the examiners, but as poor quality by the classifier. Slide *C* has clean signals and an even background to the human eye. However, the signals of the positive controls are extremely weak and not responsive to 2× reduction of the sample lysate dilution. Therefore, the QC model categorized slide *C* as poor quality

Our validation study demonstrated that the classifier is capable of producing reliable estimates of slide quality.

7 Discussion

In this study, we developed a classifier for assessing the quality of RPPA slides by using a GLM and logistic function. The GLM is a flexible generalization of ordinary least-square regression, and the logistic function computes probability, which allows for the prediction of a discrete outcome when a cutoff is chosen. The GLM is more robust than the linear regression model because it allows the magnitude of the variance of each measurement to correlate with its predictive value.

The classifier calculates the probability that a slide is of good quality. To ensure that the classifier operates with a high level of accuracy, we first fitted a GLM to all features that were potentially useful in predicting the response variable. We then tested the fit of the model after each coefficient was removed from the model during a stepwise regression process. The stepwise regression ensures that a model still adequately fits the data after variables are eliminated from the model in the iterative process. The features left in the model after AIC stepwise selection were the slopes and coefficients of variation of the positive control intensities, and the DRs and an index of the percentage of good spots for the sample intensities.

The utility of a predictive model depends on its external validity, that is, its ability to maintain accuracy when applied to data from settings that are different than those on which the model was developed.

The application of our classifier to two independent datasets that were also evaluated by expert visual examination resulted in discrepancies between the model prediction and the visual evaluation. A re-evaluation of the involved slides by the examiners increased the predictive accuracy of the classifier.

The RPPA Core Facility at The University of Texas MD Anderson Cancer Center has hybridized more than 18 000 RPPA slides each year. Because the classifier was developed and implemented in the 'SuperCurve' R package (http://bioinformatics.mdanderson.org/OOMPA/) in 2011, it has automated the RPPA production line, saved a tremendous amount of time for the RPPA experts who previously provided visual examinations of slide quality and provided high-quality data without bias to world-wide customers. These high-quality RPPA data have been applied to characterize and provide outcome predictions for studies of human colon and rectal cancer (Cancer Genome Atlas Network, 2012a), breast cancer (Cancer Genome Atlas Network, 2012b), ovarian cancer (Yang *et al*., 2013), endometrial cancer (Cancer Genome Atlas Research Network, 2013) and pan-cancers (Akbani *et al*., 2014).

All of the RPPA slides fabricated by the RPPA Core Facility are designed with samples that have five dilution steps. In order to maximize predictability, the classifier was developed based on predictive features obtained from the individual dilution steps. Therefore, the limitation of the classifier is that it is specific to the slide design. The current version of the classifier can predict the quality of an RPPA slide only if it has five dilution steps. However, the approaches used to define the classifier could be readily applied to alternative slide designs.

Acknowledgements

The authors thank Lee Ann Chastain and Rebecca I. Partida for proofreading the manuscript.

Funding

This work was supported in part by the National Institutes of Health and the National Cancer Institute through the Cancer Center Support Grant to MD Anderson Cancer Center [grant number CA016672].

Conflict of Interest: none declared.

References

Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Cont.*, **19**, 716–723.

Akbani,R. *et al.* (2014) A pan-cancer proteomic perspective on the cancer genome atlas. *Nat. Commun.*, **5**, 3887.

Cancer Genome Atlas Network. (2012a) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.

Cancer Genome Atlas Network. (2012b) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Cancer Genome Atlas Research Network *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

Carey,M.S. *et al.* (2010) Functional proteomic analysis of advanced serous ovarian cancer using reverse phase protein array: TGF-beta pathway signaling indicates response to primary chemotherapy. *Clin. Cancer Res.*, **16**, 2852–2860.

Chen,Y. *et al.* (2011) Combined Src and ER blockade impairs human breast cancer proliferation in vitro and in vivo. *Breast Cancer Res. Treat.*, **128**, 69–78.

Cohen,J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.*, **70**, 213–220.

Davies,M.A. *et al.* (2009) Integrated molecular and clinical analysis of AKT activation in metastatic melanoma. *Clin. Cancer Res.*, **15**, 7538–7546.

Dobson,A.J. and Barnett,A.G. (2008) *Introduction to Generalized Linear Models*, 3rd edn. Chapman and Hall, Boca Raton, FL.

Grote,T. *et al*. (2008) Validation of reverse phase protein array for practical screening of potential biomarkers in serum and plasma: accurate detection of CA19-9 levels in pancreatic cancer. *Proteomics*, **8**, 3051–3060.

Hardin,J. and Hilbe, J. (2007) *Generalized Linear Models and Extensions*, 2nd edn. Stata Press, College Station, TX.

Hendricks,W.A. and Robey,K.W. (1936) The sampling distribution of the coefficient of variation. *Ann. Math. Stat*., **7**, 129–132.

Liang,H. *et al*. (2012) Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Res*., **22**, 2120–2129.

Nishizuka,S. *et al*. (2003) Proteomic profiling of the NCI-60 cancer cell lines using new high-density reverse-phase lysate microarrays. *Proc. Natl. Acad. Sci. USA*., **100**, 14229–14234.

Park,E.S. *et al*. (2010) Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. *Mol. Cancer Ther*., **9**, 257–267.

Paweletz,C.P. *et al*. (2001) Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*, **20**, 1981–1989.

Tibes,R. *et al*. (2006) Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol. Cancer Ther*., **5**, 2512–2521.

Tsao,T. *et al*. (2010) Role of peroxisome proliferator-activated receptor-gamma and its coactivator DRIP205 in cellular responses to CDDO (RTA-401) in acute myelogenous leukemia. *Cancer Res*., **70**, 4949–4960.

Yang,J.Y. *et al*. (2013) Predicting time to ovarian carcinoma recurrence using protein markers. *J. Clin. Invest*., **123**, 3740–3750.