Research article

Evaluation of normalization methods for cDNA microarray data by k-NN classification

Wei Wu*1,2, Eric P Xing3, Connie Myers1, I Saira Mian1 and Mina J Bissell1

Address: 1Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, 2Dorothy P. and Richard P. Simmons Center for Interstitial Lung Disease, Division of Pulmonary, Allergy and Critical Care Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA and 3Center for Automated Learning and Discovery and Language Technology Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Email: Wei Wu* - wuw2@upmc.edu; Eric P Xing - epxing@cs.cmu.edu; Connie Myers - camyers@lbl.gov; I Saira Mian - smian@lbl.gov; Mina J Bissell - mjbissell@lbl.gov

* Corresponding author

Abstract

**Background:** Non-biological factors give rise to unwanted variations in cDNA microarray data. There are many normalization methods designed to remove such variations. However, to date there have been few published systematic evaluations of these techniques for removing variations arising from dye biases in the context of downstream, higher-order analytical tasks such as classification.

**Results:** Ten location normalization methods that adjust spatial- and/or intensity-dependent dye biases, and three scale methods that adjust scale differences were applied, individually and in combination, to five distinct, published, cancer biology-related cDNA microarray data sets. Leave-one-out cross-validation (LOOCV) classification error was employed as the quantitative end-point for assessing the effectiveness of a normalization method. In particular, a known classifier, k-nearest neighbor (k-NN), was estimated from data normalized using a given technique, and the LOOCV error rate of the ensuing model was computed. We found that k-NN classifiers are sensitive to dye biases in the data. Using NONRM and GMEDIAN as baseline methods, our results show that single-bias-removal techniques which remove either spatial-dependent dye bias (referred later as spatial effect) or intensity-dependent dye bias (referred later as intensity effect) moderately reduce LOOCV classification errors; whereas double-bias-removal techniques which remove both spatial- and intensity effect reduce LOOCV classification errors even further. Of the 41 different strategies examined, three two-step processes, IGLOESS-SFILTERW7, ISTSPLINE-SLLOESS and IGLOESS-SLLOESS, all of which removed intensity effect globally and spatial effect locally, appear to reduce LOOCV classification errors most consistently and effectively across all data sets. We also found that the investigated scale normalization methods do not reduce LOOCV classification error.

**Conclusion:** Using LOOCV error of k-NNs as the evaluation criterion, three double-bias-removal normalization strategies, IGLOESS-SFILTERW7, ISTSPLINE-SLLOESS and IGLOESS-SLLOESS, outperform other strategies for removing spatial effect, intensity effect and scale differences from cDNA microarray data. The apparent sensitivity of k-NN LOOCV classification error to dye biases suggests that this criterion provides an informative measure for evaluating normalization methods. All the computational tools used in this study were implemented using the R language for statistical computing and graphics.

Published: 26 July 2005

BMC Bioinformatics 2005, 6:191 doi:10.1186/1471-2105-6-191

Received: 17 December 2004

Accepted: 26 July 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/191

© 2005 Wu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Background

Molecular profiling technology allows for the simultaneous assaying of the abundance of tens of thousands of transcripts in a biological sample. Once these abundance values have been obtained for many samples, prevalent higher-order data analyses may include clustering, classification, feature selection, and network estimation. A variety of algorithms seeking to address these higher-order tasks have been investigated and applied, to interpret gene expression patterns and to generate biological predictions. However, the accuracy of these predictions may depend on the low-level transformations utilized to produce abundance values from raw measurements, i.e., data pre-processing may be a critical factor in determining the validity and success of downstream studies. Some key pre-processing steps for profiling data include image quantification and normalization. Several image analysis software (e.g., GenePix and SPOT) have been designed for image analysis of the spots on microarrays [1,2]. Background estimation has also been considered as an important issue in image quantification, however, evidence [2,3] showed that 'inappropriate' local background adjustment could add noise into the microarray data and thus be detrimental to the downstream studies. Background adjustment, therefore, is still an issue to be resolved. After image analysis, normalization usually needs to be performed. It is a procedure designed to minimize the unwanted variations in measurements arising from the technology, but to retain the intrinsic biological variations, and is also the focus of this work. In this study, we examined normalization in the context of a particular transcriptional profiling platform, cDNA microarrays [4-6], and the specific analytical task of classifying biological samples characterized by gene expression profiles.

In cDNA microarray-based investigations, RNA from two samples are reverse-transcribed and labeled with distinct (red and green) fluorescent dyes, then hybridized to a microarray spotted with DNA sequences (“probes”). An ensuing scanned image of the microarray is processed to yield an intensity measurement for each dye at every spot (Figure 1). If \( R \) and \( G \) are the spot-specific, quantitated, fluorescent intensities of the target and reference expression signals respectively, relative gene expression is defined as the log ratio \( M = \log_2(\frac{R}{G}) \), and average expression is the log intensity \( A = \log_2(\sqrt{RG}) \). Based on different biological assumptions and design principles, many normalization methods for cDNA microarray data have been proposed. Global normalization techniques adjust the center (e.g., mean or median) of the distribution of the log ratio \( M \) values on each microarray to a constant [1,7-9]. These methods, however, do not correct for any intensity- or spatial effect.

A variety of techniques have been proposed to remove intensity effect. A non-linear approach employs robust locally weighted regression (lowess) [10] to smooth the dependence of log ratios on intensities [4,11,12]. The basic assumption of this approach is either that the majority of genes are not differentially expressed, or that genes are influenced by random effects (i.e., the numbers of up-regulated and down-regulated genes are similar) [4,11,12]. A 'qspline' method uses a target array to adjust \( R \) and \( G \) values so that their distribution is similar for all arrays [13], but the performance of this method may depend upon the choice of the baseline array [14]. A composite method employs both external control samples and total genes on a microarray to remove intensity effect [15]. To relax critical biological assumptions, 'housekeeping-gene'-related methods first identify non-differentially-expressed genes, and then use these genes for normalization [16-18]. Semi-linear models are designed to account for the effects of print-tips (PTs), signal intensity, and
differences in gene expression levels jointly in a single model [19,20].

The removal of intensity effect at the PT level can partially remove spatial effect [4,11]. To remove spatial effect more completely, the dependence of M values on physical position can be smoothed using lowess [12], or can be eliminated using weighted mean [13] or median filter methods [17], both of which assume that differentially expressed genes are not co-localized in the neighboring spots. Since spatial- and intensity effect may be mutually dependent, a method that removes global spatial effect and global intensity effect in a single step has been proposed [21].

Whereas the above location normalization methods remove spatial- and intensity effect, scale normalization methods adjust differences in the scale of M values within and/or between microarrays. The assumption is that since the majority of genes are not differentially expressed, the scale of their M values should be constant. A robust estimate of the scale factor for scale normalization is median absolute deviation [15].

Normalization approaches seek to ensure that dye effect is removed, while biological variations are retained. Spatial and intensity effect and scale effect arise from printing, hybridization, scanning, or other technical factors, and can mask the signals arising from genuine biological variations in gene expression. Visual aids used to assess the effectiveness of normalization methods [11,13,15,21] include scatter plots of log ratio (M) versus average log intensity (A) ("MA plots"). Spatial plots are a color-coded representation of each spot on a microarray that depicts M values, or a quality (e.g., shape, size) measure of some test statistic. These two types of diagnostic plots [4,21] suggest that raw M values are often biased estimates of relative expression and that the dye intensities per spot need to be adjusted. Quantitative criteria used to assess the robustness of normalization methods in removing dye effect include (i) rank variations of spot intensity in non-normalized versus normalized data [9,22], and (ii) correlation [16,21], variance [8,13], or error [18,22] of the normalized M values in replicated data.

To ensure that biological variations are retained after normalization, several functional criteria have been employed. Prevailing approaches determine the ability to predict a fixed number of differentially expressed genes in real or simulated data using quantitative measures based on t-statistics [4,11,13,21], adjusted p-values [11], and false-discovery rates [23]. However, there is uncertainty associated with these measures, and the true number of differentially expressed genes is unknown. Spike-in data have been used to assess normalization approaches for Affymetrix GeneChip data [14,24,25]. However, external control samples are not widely used for evaluation of normalization methods for cDNA microarrays.

In this paper, we evaluated normalization methods for cDNA microarray data using the k-NN LOOCV classification error (of biological samples characterized by the gene expression profiles), an alternative quantitative functional measure that is relatively unambiguous, objective and readily computed. We used k-NN classifiers because (i) their sensitivity enables us to discriminate between, and hence evaluate normalization techniques, (ii) they are readily available, (iii) they perform well in practice, and (iv) their non-parametric nature means that assumptions about how the data are distributed have little influence on classification performance. Since the primary aim of our evaluation of normalization methods was to assist practitioners in choosing effective data pre-processing schemes, we did not consider factors that may influence classification performance, such as feature selection and distance metrics. We investigated a wide spectrum of well-known and widely available normalization techniques: ten location normalization methods that adjust spatial effect and/or intensity effect (Table 1), and three scale methods that adjust scale differences (Table 3). We applied these methods, individually and in combination (41 strategies in all, Tables 1, 2, 3), to five diverse, published, cancer biology-related cDNA microarray data sets (Table 4), and we generated data sets with spatial effect, intensity effect and scale differences removed to varying degrees. Computing the LOOCV classification error of k-NNs estimated from these multi- and two-class data sets allowed us to investigate which and how much of the dye effect are removed by the 41 strategies.

**Results**

### Spatial- and intensity-dependent normalization

**Diagnostic plots**

We used diagnostic plots to examine the ability of different location normalization methods to remove spatial- and/or intensity effect (Tables 1 and 2). Figure 2 shows spatial plots for two specific LYMPHOMA microarrays normalized with four approaches designed to correct spatial effect (SLOESS, SLFILTERW3, SLFILTERW7, IGSGLOESS). The non-normalized M values (NONRM) for microarray "5850" display global spatial effect (left-to-right, green-to-red pattern) whereas those for microarray "5938" exhibit local spatial effect (top-to-bottom, green-to-red pattern) in each PT group. Removal of spatial effect should result in a "random" red and green pattern of M values. SLOESS and SLFILTERW7 exhibit similar dye bias-removal abilities in that they both remove global spatial effect more effectively than local spatial effect. SLFILTERW3 removes both global and local dye effect effectively, largely because it uses a median filter of a small window size (3 x 3 spots) for normalization. IGSGLOESS...
removes most, but not all, global and local spatial effect (a strip of red spots on the right side of "5850" and on the bottom of the PT groups in the first row of "5938" remain). IGSGLOESS may not be as effective at removing dye effect as expected because, as the developers indicate, lowess curve construction uses the standardized spatial variables (rloc, cloc), which may not be appropriate for location variables [21].

Figure 3 shows intensity-dependent MA plots for one specific LYMHPMA microarray overlaid with one lowess curve (left) or one lowess curve per print tip group (right) using six methods designed to correct intensity effect (IGLOESS, ILOESS, ISTSPINE, QSPLINEG, QSPLINER, IGSGLOESS). For non-normalized M values (NONRM), the curvature in the MA plot indicates the presence of intensity effect at the array (left) and PT (right) level. All six methods remove global intensity effect completely (flat lowess curves, left), but only ILOESS and IGSGLOESS remove local intensity effect thoroughly (right).

Visual inspection of the diagnostic plots in Figures 2 and 3 suggest that SLLFILTERW3 is an effective method for removing both global and local spatial effect, whereas ILOESS is good at removing intensity effect.

### k-NN LOOCV Classification error

For a functional, quantitative evaluation of location normalization methods, we first computed k-NN LOOCV classification error rates for data normalized using these methods individually and/or in combination. Then for each data set, we ranked the normalization methods based on their LOOCV classification error rates. The smaller the LOOCV classification error rate, the lower the rank of the normalization strategy. In order to assess whether normalization is beneficial (or not), we also
computed the following quantity for a normalization method in each data set:

$\text{IMPROVEMENT} = \frac{(\text{ErrorRate(NONRM)} - \text{ErrorRate(Method)})}{\text{ErrorRate(NONRM)}} \times 100\%$

where ErrorRate(NONRM) is the error rate of NONRM, and ErrorRate(Method) is the error rate of the method. Tables 5 and 6 give results for five data sets (Table 4) and 23 location methods designed to remove spatial- and/or intensity effect (Tables 1 and 2). Figures 4 and 5 are alternative, visual representations of the classification "Error Rate" and "Rank" in Table 5.
Single-bias-removal methods
These strategies can be classified into two categories, spatial-dependent and intensity-dependent normalization methods. Three spatial-dependent normalization methods (SLOESS, SLFILTERW3, SLFILTERW7) reduce k-NN LOOCV classification error rates to a similar extent (Tables 5 and 6) and have almost identical ranks (Figure 5), despite the fact that their abilities to remove spatial effect are quite different (Figure 2). Since both SLOESS and SLFILTERW7 fail to remove local spatial patterns effectively (Figure 2, rows 2 and 4), SLFILTERW3 may be too aggressive in removing "dye effect" (Figure 2, row 3). However, the three intensity-dependent methods (IGLOESS, IILLOESS, ISTSPILINE) reduce k-NN LOOCV classification error rates to different degrees. The k-NN LOOCV classification error rate and rank of IGLOESS are similar to those of the three spatial-dependent methods (SLOESS, SLFILTERW3, SLFILTERW7) (Figure 5), whereas IILLOESS, which removes intensity effect more completely than IGLOESS, has smaller k-NN LOOCV classification error rates than IGLOESS in all five data sets. ISTSPILINE, which uses a rank invariant set for normalization, is also better than IGLOESS in all five data sets (Figure 5).

In all five data sets, except for LYMPHOMA (SLOESS), the single-bias-removal normalization methods consistently yield smaller LOOCV classification error rates than no-bias-removal methods, NONRM and GMEDIAN (which only sets the median of the distribution of M values to zero). The greatest benefit, an IMPROVEMENT of 56%, is seen with GASTRIC CARCINOMA (SLOESS, ISTSPILINE) (Table 6).

Double-bias-removal methods
IGSGLOESS removes both spatial- and intensity-effect in one step, whereas the remaining seven approaches are two-step strategies consisting of single-bias-removal methods applied sequentially (first a method to remove intensity effect, followed by a method to remove spatial effect).

In general, double-bias-removal methods have smaller k-NN LOOCV classification error rates and bigger IMPROVEMENT than single-bias-removal methods, and
Table 4: The multi-class, cancer-biology related transcriptional profiling data sets analyzed in this work. For each of the five published studies, the fluorescent intensities, microarray images, and associated information were downloaded from the URLs indicated. The studies refer to data sets produced after application of all pre-normalization data processing, location/scale normalization, and post-normalization data processing steps. The abbreviations are as follows, Microarrays: number of cDNA microarrays; Probes: number of probes; K: total number of categories to which a sample could be assigned; Samples and Class: number of samples in the specified pre-defined category; Configuration: configuration of a microarray using the convention described in Figure 1.

| Data set name       | Description                                                                 |
|---------------------|-----------------------------------------------------------------------------|
| LIVER CANCER [46]   | Microarrays: 181; Probes: 6,605; K = 2                                     |
|                     | Samples and Class: 76 normal; 105 tumor                                     |
|                     | Configuration: (ngr = 8, ngc = 4, nsr = 27, nsc = 28)                       |
|                     | http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=107  |
| LYMHPOMA [47]       | Microarrays: 81; Probes: 6,850; K = 3                                      |
|                     | Samples and Class: 29 normal, 43 diffuse large B-cell lymphoma (DLBCL); 9 follicular lymphoma (FL) |
|                     | Configuration: (ngr = 4, ngc = 4, nsr = 24, nsc = 24); (ngr = 8, ngc = 4, nsr = 24, nsc = 24) |
|                     | http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=79   |
| RENAL CELL CARCINOMA [48] | Microarrays: 38; Probes: 13,608; K = 4                                    |
|                     | Samples and Class: 3 normal; 26 clear cell carcinoma (CCC); 5 granular cell carcinoma (GCC); 4 papillary carcinoma (PC) |
|                     | Configuration: (ngr = 8, ngc = 4, nsr = 27, nsc = 28)                       |
|                     | http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=210  |
| GASTRIC CARCINOMA [49] | Microarrays: 130; Probes: 15,541; K = 2                                  |
|                     | Samples and Class: 28 normal; 102 tumor                                     |
|                     | Configuration: (ngr = 12, ngc = 4, nsr = 30, nsc = 32); (ngr = 12, ngc = 4, nsr = 29, nsc = 32); (ngr = 12, ngc = 4, nsr = 30, nsc = 30) |
|                     | http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=232  |
| LUNG CANCER [50]    | Microarrays: 60; Probes: 20,601; K = 5                                     |
|                     | Samples and Class: 6 normal; 35 adenocarcinoma (AC); 11 squamous cell carcinoma (SCC); 4 large cell lung cancer (LCLC); 4 small cell lung cancer (SCLC) |
|                     | Configuration: (ngr = 8, ngc = 4, nsr = 27, nsc = 28)                       |
|                     | http://genome-www5.stanford.edu/cgi-bin/publication/viewPublication.pl?pub_no=100  |

The ranks of the SLFILTERW3-related approaches (IGLOESS-SLFILTERW3, ISTDPLINE-SLFILTERW3, QSPLINEG-SLFILTERW3, QSPLINER-SLFILTERW3) are higher than their SLFILTERW7 counterparts (Figure 5), suggesting that a window size of 7 × 7 is more preferable than that of 3 × 3. A smaller window size may over normalize the data, and thus conceal real biological variations.

Compared to the two-step approaches, the rank of the one-step method, IGSLOESS, is higher than IGLOESS-SLFILTERW7 and ISTDPLINE-SLFLOESS (yet lower than IGLOESS-SLFILTERW3 and ISTDPLINE-SLFILTERW3). This indicates that the one-step IGSLOESS has no apparent advantage over the two-step bias-removal strategies.

Overall, the classification performances of data normalized using the double-bias-removal methods are better than that of NONRM, and the benefits (IMPROVEMENT) of doing so range from 21% in the case of LUNG CANCER.
IGS-LOESS) to 100% in GASTRIC CARCINOMA (IGS-LOESS) (Table 6).

Qspline-related approaches

Unlike the location normalization methods discussed above, qspline-related approaches require a target array. QSPLINEG and QSPLINER are single-bias-removal techniques and use G and R respectively as the target array. The reduction in k-NN LOOCV classification error rates for these methods is quite significant compared to the other single-bias-removal methods. However, it is noticeable that although QSPLINEG and QSPLINER produce similar results in almost all data sets, their results are different in LYMPHOMA (Figures 4 and 5). In addition, when QSPLINEG or QSPLINER is combined with one of the three spatial-dependent methods, the rank of the resulting double-bias-removal technique is different from that of its counterpart technique (Figure 5). These results suggest that, similar to other baseline array-based normalization methods [14], the performances of the qspline-related methods may also depend on the choice of the target array.

Overall, the classification performance of data normalized using the qspline-related methods is better than NONRM by IMPROVEMENT values of 9% in LUNG CANCER (QSPLINEG-SLFILTERW3) and of 100% in GASTRIC CARCINOMA (QSPLINEG, QSPLINER). None of these qspline-related methods, however, outperforms the IGOESS-SLFILTERW7 (Table 6).

Scale normalization

Figure 6 shows boxplots of the distribution of non-normalized M values for microarrays in the five studies. Scale effect is more apparent between (right) rather than within (left) microarrays in a study. The LYMPHOMA data set shows considerable variations in box size and whisker length both within and between microarrays.
Tables 7 and 8 and Figure 7 show LOOCV classification error rates, ranks and IMPROVEMENT for the $k$-NN classifiers estimated using 3 scale normalization methods combined with other spatial- and/or intensity-dependent normalization methods (18 strategies in all). For data normalized first with spatial- and/or intensity-dependent methods, little or no reduction in LOOCV classification error rates was observed when within-microarray scale normalization ($W_{\text{SCALE}}$) was applied later. However, when between-microarray scale normalization ($B_{\text{SCALE}}$) was used alone, or when both scale normalization techniques were used sequentially ($W_{\text{SCALE}}$), there was an increase in both median and upper quantile ranks (Figure 7), suggesting that $B_{\text{SCALE}}$ should not be applied on the studied data sets. With regard to our running example, the LYMPHOMA data set, scale normalization has no apparent beneficial effect on classification performance.

Discussion
This computational investigation employed two types of visual diagnostic plots and $k$-NN LOOCV classification error rates to evaluate a broad suite of known normalization strategies. These analyses were applied to cDNA microarray data from five published cancer studies. Since all these data sets were acquired using GenePix image analysis software and a recent study showed that background adjustment using GenePix can increase variability of microarray data and compromise downstream data analyses [3], we used foreground intensity values of the probes without background adjustment in this work. The normalization approaches examined are based on a variety of different techniques and implementations that are readily available and accessible.
Our results show that the LOOCV classification error of \textit{k}-NN classifiers depends on how much of spatial- and intensity effect can be removed by a normalization strategy. Overall, the single-bias-removal location approaches perform better than GMEDIAN and NONRM, while the double-bias-removal location strategies perform better than the single-bias-removal location approaches. Of the twenty-three location normalization techniques investigated, three two-step processes (IGLOESS-SFILTERW7, ISTSPLINE-SLOESS and IGLOESS-SLOESS), all of which removes intensity effect at the global level and spatial effect at the local level, appear to be the most effective at reducing LOOCV classification error. However, removing spatial- or intensity effect alone is not sufficient for reducing LOOCV classification error (see below).

A recent review of normalization methods [26] raised the concern that removing spatial effect (SLOESS and the related methods) may add additional noise to normalized data, and suggested that a safe alternative was removing only intensity effect at the local level (ILLOESS) [26]. Our results show that, although the classification performance of data normalized with SLOESS alone can be worse than non-normalized data as in the case of the LYMPHOMA data set, when SLOESS is combined with another intensity-dependent approach (IGLOESS, ILLOESS, ISTSPLINE, QSPLINEG, or QSPLINER), there is considerable improvement over NONRM, with IMPROVEMENT ranging from 23% in LIVER CANCER (QSPLINER-SLOESS) to 78% in GASTRIC CARCINOMA (QsPLINER-SLOESS, QsPLINEG-SLOESS). Thus, removing both spatial- and intensity effect is beneficial for the downstream analytical task of classification. Another study compared various \textit{lowess}-based single-bias-removal intensity normalization approaches, and found that ILLOESS may not significantly improve the results compared to IGLOESS [27]. Our results show that the benefits (IMPROVEMENT) of IGLOESS over NONRM.
range from 5% in LUNG CANCER to 44% in RENAL CELL CARCINOMA; while that the benefits (IMPROVEMENT) of ILLOESS over NONRM range from 23% in RENAL CELL CARCINOMA to 46% in LIVER CANCER. Therefore, ILLOESS performs better than ILOESS in our study. However, as a single-bias-removal approach, ILLOESS still fail to outperform ILOESS-SLLOESS, ISTSPLINE-SLOESS and ILOESS-SLLOESS, which are the best overall methods and whose IMPROVEMENT values over NONRM range from 40% in LUNG CANCER to 58% in LYMPHOMA for ILOESS-SLFILTERW7, from 33% in GASTRIC CARCINOMA to 62% in LYMPHOMA for ISTSPLINE-SLOESS and from 33% in LUNG CANCER to 56% in GASTRIC CARCINOMA for ILOESS-SLLOESS (Table 6).

A previous study employed k-NN classification of diluted samples to assess a small number of global linear methods for normalization [28]. The study presented here is more comprehensive, both in terms of the range of data sets and the diversity of normalization techniques. Our results indicate that the k-NN LOOCV classification error rate of real biological samples provides an informative functional quantitative measure that can be used to evaluate normalization approaches.

Differences in scale between microarrays can arise from both unwanted technical factors (differences in experimental reagents, equipment, personnel, and so on), as well as from genuine biological variations. The scale normalization techniques applied here aim to remove unwanted technical factors, and assume the existence of little biological variations between samples. For the five studied data sets, scale normalization of non- or location-normalized data do not result in an overall reduction in LOOCV classification error. Indeed, two between-microarray normalization methods (BSCALE, WBSCALE) result in an overall increase in LOOCV classification error (poorer performance, Figure 7). These results suggest that in the examined cancer-related data sets, there can be con-

---

Table 6: IMPROVEMENT of location normalization methods. IMPROVEMENT is defined (in the Results) based on improvement of LOOCV classification error rate of a given normalization method over that of NONRM. The methods are arranged in the same order as those in Table 5. For a given data set, the biggest IMPROVEMENT(s) is shown in bold. The methods and data sets are described in Tables 1, 2 and 4, respectively.

| Location Normalization method | IMPROVEMENT (%) | IMPROVEMENT (%) | IMPROVEMENT (%) | IMPROVEMENT (%) | IMPROVEMENT (%) |
|------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| NONRM                        | 0               | 0               | 0               | 0               | 0               | 0-0             |
| GMEDIAN                      | 19              | 7               | 33              | 22              | 5               | 5-33            |
| SLLOESS                      | 33              | -2              | 44              | 45              | 0               | 0-45            |
| SLFILTERW3                   | 23              | 19              | 44              | 45              | 0               | 0-45            |
| SLFILTERW7                   | 29              | 5               | 44              | 34              | 9               | 5-44            |
| ILOESS                       | 34              | 30              | 44              | 33              | 5               | 5-44            |
| ILOESS-SLLOESS               | 46              | 42              | 44              | 33              | 23              | 23-46           |
| ILOESS-SFILTERW7             | 36              | 33              | 52              | 56              | 7               | 7-56            |
| ILOESS-SFILTERW3             | 33              | 51              | 44              | 100             | 21              | 21-100          |
| ILOESS-SFILTERW7             | 44              | 56              | 50              | 56              | 33              | 33-56           |
| ILOESS-SFILTERW3             | 48              | 58              | 44              | 44              | 26              | 26-58           |
| ILOESS-SFILTERW7             | 44              | 58              | 50              | 56              | 40              | 40-58           |
| QSPLINEG-SFILTERW7           | 40              | 62              | 50              | 33              | 47              | 33-62           |
| QSPLINEG-SFILTERW3           | 22              | 48              | 61              | 34              | 42              | 22-61           |
| QSPLINEG-SFILTERW7           | 42              | 52              | 44              | 34              | 42              | 34-52           |
| QSPLINEG-LSFILTERW7          | 22              | 28              | 59              | 100             | 23              | 22-100          |
| QSPLINER                     | 18              | 54              | 59              | 100             | 23              | 18-100          |
| QSPLINER-LSFILTERW7          | 32              | 26              | 50              | 78              | 37              | 26-78           |
| QSPLINER-LSFILTERW3          | 29              | 30              | 27              | 78              | 37              | 12-78           |
| QSPLINER-LSFILTERW7          | 26              | 28              | 55              | 78              | 37              | 26-78           |
| QSPLINER-LSLOESS             | 23              | 61              | 56              | 78              | 37              | 23-78           |
| QSPLINER-LSFILTERW3          | 23              | 58              | 39              | 78              | 9               | 9-78            |
| QSPLINER-LSFILTERW7          | 16              | 56              | 50              | 67              | 23              | 16-67           |

---

BMC Bioinformatics 2005, 6:191 http://www.biomedcentral.com/1471-2105/6/191
Bar plots of leave-one-out cross-validation error rates for k-NNs in Table 5. The classifiers were estimated from five data sets (Table 4) either without normalization (NONRM) or normalized using twenty-three normalization techniques that remove spatial- and/or intensity effect to varying degrees (Tables 1 and 2). In each plot, the normalization methods are arranged in the following order: (A) Methods that remove no dye bias (GMEDIAN), or a single dye bias (SLOESS, SLFILTERW3, SLFILTERW7, IGLOESS, ILLOESS, ISTSPLINE). (B) Methods that remove two dye biases (IGSGLOESS, IGLOESS-SLLOESS, ILLOESS-SLLOESS, IGLOESS-SLFILTERW3, IGLOESS-SLFILTERW7, ISTSPLINE-SLOESS, ISTSPLINE-SLFILTERW3, ISTSPLINE-SLFILTERW7). (C) Qspline-related methods (QSPLINEG, QSPLINER, QSPLINEG-SLLOESS, QSPLINEG-SLFILTERW3, QSPLINEG-SLFILTERW7, QSPLINER-SLOESS, QSPLINER-SLFILTERW3, QSPLINER-SLFILTERW7).
### Rank summary for location normalization methods

| Location Normalization Method | Median Rank | Upper Quantile Rank |
|-------------------------------|-------------|---------------------|
| Nonorm                        | 24          | 24                  |
| Gmedian                       | 21          | 22                  |
| sLloess                       | 16.5        | 22                  |
| sLfilterW3                    | 16.5        | 20                  |
| sLfilterW7                    | 16.5        | 17.5                |
| iGloess                       | 16.5        | 20                  |
| iLloess                       | 13          | 16.5                |
| iSTSpline                     | 10          | 14                  |
| iGSGloess                     | 10          | 15                  |
| iGloess–sLloess               | 8.5         | 10                  |
| iLloess–sLloess               | 10          | 15                  |
| iGloess–sLfilterW3            | 11          | 19.5                |
| iGloess–sLfilterW7            | 4           | 10                  |
| iSTSpline–sLloess             | 6           | 10                  |
| iSTSpline–sLfilterW3          | 12          | 17.5                |
| iSTSpline–sLfilterW7          | 9           | 16.5                |
| QsplineG                      | 12.5        | 17.5                |
| QsplineR                      | 8           | 12.5                |
| QsplineG–sLloess              | 10          | 11                  |
| QsplineG–sLfilterW3           | 15.5        | 16                  |
| QsplineG–sLfilterW7           | 6           | 14                  |
| QsplineR–sLloess              | 6           | 7.5                 |
| QsplineR–sLfilterW3           | 16          | 17.5                |
| QsplineR–sLfilterW7           | 10          | 12.5                |

**Figure 5**

**Rank summary for location normalization methods.** The median and upper quantile ranks of each method are defined as the median and upper quantile values of the ranks of each method across all five data sets (see Table 5, "Ranks"). The bar plots present a visual depiction of the results in the table. (Median ranks are shown in pink; upper quantile ranks are shown in blue.)
Boxplots of the distributions of non-normalized $M$ values for microarrays in the five studies

In each boxplot, the box depicts the main body of the data and the whiskers show extreme values. The variability is indicated by the size of the box and the length of the whiskers ($marray/marraymaBoxplot(y="maM")$). Each panel in the left-hand column shows results for $M$ values at the local level of a microarray chosen at random from a given data set. The bars are color-coded by PT group. Each panel in the right-hand column shows results for $M$ values at the global level for 50 microarrays chosen at random from a given data set (the total number of microarrays in RENAL CELL CARCINOMA is 38). Each row corresponds to a particular study.
siderable genuine biological variations (which is plausible because genomic aberrations found in cancer cells [29,30] may alter the number and nature of expressed genes compared to normal cells), and that these variations are masked by the applied scale normalization. The data sets considered here do not contain replicated data, so it is difficult to ascertain how much of the scale effect result from unwanted technical factors. Scale normalization may be warranted in situations where technical differences can be discerned by examination of the replicated data and genuine biological variations are known or believed to exist. In such cases, scale normalization using external control samples may be more useful than the total gene approaches.

While our empirical analyses are thoroughgoing in terms of both normalization procedures and test data sets, we acknowledge that there are two caveats in this study that deserve attention and further investigation. First, we employed the LOOCV classification error as a functional measure to assess normalization methods. In principle, LOOCV provides an almost unbiased estimate of the generalization ability of a classifier [31], especially when the number of the available training samples is severely limited (as in the case of LYMPHOMA and RENAL CELL CARCINOMA in Table 4), and is thus highly desirable for model selection or other relevant algorithm evaluation [32,33]. However, it is also known that the LOOCV error estimator may have high variance in some situations [34,35], which could in turn affect the accuracy of the rankings of the normalization methods. Empirically, however, we found that the LOOCV errors we obtained from various round of classification are quite stable, therefore we believe that our estimation is in practice reliable and suitable for ranking. Nevertheless, error estima-

| Location, Scale Normalization method | LIVER CANCER | LYMPHOMA | RENAL CELL CARCINOMA | GASTRIC CARCINOMA | LUNG CANCER |
|-------------------------------------|-------------|-----------|----------------------|------------------|-------------|
| Error Rate Rank                     | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank |
| NONRM                               | 0.202 19.5 0.266 18.5 0.237 24 | 0.0347 14.5 0.359 19.5 |
| NONRM, WSCALE                       | 0.185 14.5 0.303 23.5 0.211 22.5 | 0.0270 12 0.350 17.5 |
| NONRM, BSCALE                       | 0.227 23.5 0.272 20.5 0.132 8 | 0.0615 22 0.425 24 |
| NONRM, WBScale                      | 0.227 23.5 0.303 23.5 0.132 8 | 0.0462 21 0.392 23 |
| GASTRIC CARCINOMA                   |             |           |                     |                  |             |
| Error Rate Rank                     | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank |
| SLLOESS                              | 0.136 9 0.272 20.5 0.132 8 | 0.0154 3 0.350 17.5 |
| SLLOESS, WSCALE                     | 0.127 7 0.266 18.5 0.132 8 | 0.0193 6 0.342 15.5 |
| SLLOESS, BSCALE                     | 0.202 19.5 0.260 17 0.145 14.5 | 0.1000 24 0.300 11.5 |
| SLLOESS, WBScale                     | 0.191 16 0.284 22 | 0.119 2 | 0.0846 23 0.283 9.5 |
| GLOESS                               | 0.133 8 0.186 15 0.132 8 | 0.0231 9.5 0.342 15.5 |
| GLOESS, WSCALE                      | 0.141 10 0.148 11 0.132 8 | 0.0193 6 0.317 13.5 |
| GLOESS, BSCALE                      | 0.215 22 0.216 16 0.158 18.5 | 0.0308 13 0.375 21 |
| GLOESS, WBScale                      | 0.202 19.5 0.179 14 | 0.158 18.5 | 0.0231 9.5 0.383 22 |
| LUNG CANCER                          |             |           |                     |                  |             |
| Error Rate Rank                     | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank | Error Rate Rank |
| ILOESS                               | 0.110 3.5 0.154 12 0.132 8 | 0.0231 9.5 0.275 7 |
| ILOESS, WSCALE                      | 0.116 6 0.161 13 0.145 14.5 | 0.0231 9.5 0.300 11.5 |
| ILOESS, BSCALE                      | 0.193 17 0.111 3 | 0.145 14.5 | 0.0385 17.5 0.359 19.5 |
| ILOESS, WBScale                      | 0.202 19.5 0.105 1 | 0.145 14.5 | 0.0424 20 0.317 13.5 |
| ILOESS-SLLOESS                       | 0.105 1.5 0.111 3 | 0.132 8 | 0.0193 6 0.267 4.5 |
| ILOESS-SLLOESS, WSCALE              | 0.105 1.5 0.136 9.5 0.132 8 | 0.00769 1 | 0.267 4.5 |
| ILOESS-SLLOESS, BSCALE              | 0.160 12 0.130 8 | 0.211 22.5 | 0.0385 17.5 0.275 7 |
| ILOESS-SLLOESS, WBScale              | 0.152 11 0.123 6 | 0.158 18.5 | 0.0347 14.5 0.259 3 |
| ILOESS-SLFLFILTERW7                 | 0.113 5 0.111 3 | 0.119 2 | 0.0154 3 0.217 2 |
| ILOESS-SLFLFILTERW7, WSCALE         | 0.110 3.5 0.124 7 | 0.119 2 | 0.0154 3 | 0.209 1 |
| ILOESS-SLFLFILTERW7, BSCALE         | 0.180 13 0.117 5 | 0.171 21 | 0.0385 17.5 0.275 7 |
| ILOESS-SLFLFILTERW7, WBScale        | 0.185 14.5 0.136 9.5 0.158 18.5 | 0.0385 17.5 0.283 9.5 |
tors that have shown to have low variance (e.g., bootstrapping and $k$-fold cross-validation [34,35]) are worth further investigation in the future.

The second caveat of this work is that normalization methods were evaluated using $k$-NN classification without the aid of auxiliary techniques, such as feature selection. The reasons we did not employ feature selection, but rather used all the probes that are present in the majority of the microarrays for classification are as follow:
i) We believe that the influence of the dye effect (which usually affect a large number of the probes) on the downstream data analysis can be better and more consistently reflected when a large number of the probes are examined. As such, using all valid probes for training a classifier can best reflect the effectiveness of the normalization methods, whereas using subsets of the probes may generate inconsistent results due to the heterogeneous nature of the dye effect across microarrays. ii) We also included low intensity probes in the analyses. Although this may add noise and therefore could compromise the absolute classification performance of the examined normalization methods, we nevertheless think that these probes should not be excluded because reducing variability in low intensity probes is by itself an important objective of normalization methods. That is, a good normalization approach should be able to reduce variability in both low intensity- and high intensity probes effectively. And iii) we are aware that $k$-NNs without feature selection may add variability to the classification results, however, $k$-NN classification is also appealing in that it is simple and requires no data pre-processing or assumption on data distribution. In addition, $k$-NN classifiers have been widely used in many classification tasks including high-dimensional problems arising from image and text data [36].

| Location, Scale Normalization method | IMPROVEMENT (%, LIVER CANCER) | IMPROVEMENT (%, LYMPHOMA) | IMPROVEMENT (%, RENAL CELL CARCINOMA) | IMPROVEMENT (%, GASTRIC CARCINOMA) | IMPROVEMENT (%, LUNG CANCER) |
|-------------------------------------|-------------------------------|-----------------------------|---------------------------------------|-----------------------------------|-------------------------------|
| NONRM                              | 0                             | 0                           | 0                                     | 0                                 | 0                             |
| NONRM, WSCALE                      | 8                             | -13                         | 11                                    | 22                                | 3                             |
| NONRM, BSCALE                      | -12                           | -2                          | 44                                    | -77                               | -18                           |
| NONRM, WBSCALE                     | -12                           | -13                         | 44                                    | -33                               | -9                            |
| SLOESS                             | 33                            | -2                          | 44                                    | 56                                | 3                             |
| SLOESS, WSCALE                     | 37                            | 0                           | 44                                    | 44                                | 5                             |
| SLOESS, BSCALE                     | 0                             | 2                           | 39                                    | -188                              | 16                            |
| SLOESS, WBSCALE                    | 5                             | -7                          | 50                                    | -144                              | 21                            |
| IGLOESS                            | 34                            | 30                          | 44                                    | 33                                | 5                             |
| IGLOESS, WSCALE                    | 30                            | 44                          | 44                                    | 44                                | 12                            |
| IGLOESS, BSCALE                    | -6                            | 19                          | 33                                    | 11                                | -4                            |
| IGLOESS, WBSCALE                   | 0                             | 33                          | 33                                    | 33                                | -7                            |
| ILOESS                             | 45                            | 42                          | 44                                    | 33                                | 23                            |
| ILOESS, WSCALE                     | 43                            | 39                          | 39                                    | 33                                | 16                            |
| ILOESS, BSCALE                     | 4                             | 58                          | 39                                    | -11                               | 0                             |
| ILOESS, WBSCALE                    | 0                             | 61                          | 39                                    | -22                               | 12                            |
| ILOESS-SLOESS                      | 48                            | 58                          | 44                                    | 44                                | 26                            |
| ILOESS-SLOESS, WSCALE              | 48                            | 49                          | 44                                    | 78                                | 26                            |
| ILOESS-SLOESS, BSCALE              | 21                            | 51                          | 11                                    | -11                               | 23                            |
| ILOESS-SLOESS, WBSCALE             | 25                            | 54                          | 33                                    | 0                                 | 28                            |
| IGLOESS-SL FILTER W7               | 44                            | 58                          | 50                                    | 56                                | 40                            |
| IGLOESS-SL FILTER W7, WSCALE       | 46                            | 53                          | 50                                    | 56                                | 42                            |
| IGLOESS-SL FILTER W7, BSCALE       | 11                            | 56                          | 28                                    | -11                               | 23                            |
| IGLOESS-SL FILTER W7, WBSCALE      | 8                             | 49                          | 33                                    | -11                               | 21                            |
Rank summary for scale normalization methods

The median ranks and upper quantile ranks are defined as described in Figure 5. The bar plots present a visual depiction of the results in the table. (Mean ranks are shown in pink; median ranks are shown in blue.) In each plot, normalization strategies are arranged in the following order: a location normalization method, a location normalization method followed by WSCALE (+WSCALE), a location normalization method followed by BSCALE (+BSCALE), a location normalization method followed by WBSCALE (+WBSCALE).

| Location, Scale normalization method | Median Rank | Upper Quantile Rank |
|-------------------------------------|-------------|---------------------|
| Nonorm                              | 19.5        | 19.5                |
| Nonorm, Wscale                      | 17.5        | 22.5                |
| Nonorm, Bscale                      | 22          | 23.5                |
| Nonorm, WBscale                     | 23          | 23.5                |
| sLloess                             | 9           | 17.5                |
| sLloess, Wscale                     | 8           | 15.5                |
| sLloess, Bscale                     | 17          | 19.5                |
| sLloess, WBscale                    | 16          | 22                  |
| iGloess                             | 9.5         | 15                  |
| iGloess, Wscale                     | 10          | 11                  |
| iGloess, Bscale                     | 18.5        | 21                  |
| iGloess, WBscale                    | 18.5        | 19.5                |
| iLloess                             | 8           | 9.5                 |
| iLloess, Wscale                     | 11.5        | 13                  |
| iLloess, Bscale                     | 17          | 17.5                |
| iLloess, WBscale                    | 14.5        | 19.5                |
| iLloess–sLloess                     | 4.5         | 6                   |
| iLloess–sLloess, Wscale             | 4.5         | 8                   |
| iLloess–sLloess, Bscale             | 12          | 17.5                |
| iLloess–sLloess, WBscale            | 11          | 14.5                |
| iGloess–sLfilterW7                  | 3           | 3                   |
| iGloess–sLfilterW7, Wscale          | 3           | 3.5                 |
| iGloess–sLfilterW7, Bscale          | 13          | 17.5                |
| iGloess–sLfilterW7, WBscale         | 14.5        | 17.5                |
Due to the above two caveats, the relative rankings of the investigated normalization strategies can hardly be obtained accurately in this work. For example, our results show that IGLOESS-SLFILTERW7, ISTEPPLINE-SLOESS and IGLOESS-SLLOESS reduce LOOCV classification errors most consistently and effectively across all five data sets. It is difficult, however, to determine further which of these three strategies is the best, because small differences in classification results can either arise from inherent differences in these approaches and/or from the variability introduced by the LOOCV error estimator and less optimal k-NN classifiers. Moreover, our results should not be taken as a warrant of directly using baseline methods, such as k-NNs without feature selection, for high-dimensional classification tasks. More investigations are needed to understand the interplay between normalization (which improves data quality) and feature selection (which improves the classifier by throwing away non-informative data) to ascertain normalization strategies to produce an optimal classifier.

Conclusion

Using LOOCV error of k-NNs as the evaluation criterion, we assessed a variety of normalization methods that remove spatial effect, intensity effect and scale differences from cDNA microarray data. Overall, the single-bias-removal location approaches (which remove either spatial- or intensity effect from the data) perform better than GMEDIAN and NONRM, while the double-bias-removal location strategies (which remove both spatial- and intensity effect) perform better than the single-bias-removal location approaches. Of the 41 different strategies examined, IGLOESS-SLFILTERW7, ISTEPPLINE-SLOESS and IGLOESS-SLBSSCALE, all of which are two-step approaches and remove both intensity effect at the global level and spatial effect at the local level, appear to be the most effective at reducing LOOCV classification error. The investigated scale normalization methods do not have beneficial effect on classification performance. These results also indicate that spatial- and intensity effect do have profound impact on downstream data analyses, such as classification, and that removing these effects can improve the quality of such analyses.

Methods

Extant data sets and software

Table 4 summarizes relevant information on the cDNA microarray data sets from the Stanford Microarray Database (SMD) reexamined here. These data sets were selected because the published studies assayed samples from distinct cancers, the profiling experiments were performed at different times, four out of the five data sets were produced by different investigators, and the microarrays used were printed with different probes on different occasions. The LYMPHOMA study has been used as the illustrative, running example.

A variety of computational tools for manipulating, analyzing, and visualizing microarray data are available. These include open source implementations based on the R language for statistical computing http://www.r-project.org[37] such as the Bioconductor http://www.bioconductor.org, MAANOVA http://www.jax.org/staff/churchill/labsite/software/download.html. tRMA http://www.pi.csiro.au/gena/tRMA, and braju http://www.maths.lth.se/help/R/aroma packages. Standard R and Bioconductor packages and functions were used apart from one normalization method found in MAANOVA (joint removal of both spatial- and intensity effect at the global microarray level, IGBSSCALE) and two normalization methods found in tRMA (removal of spatial effect at the local PT level, SLFILTERW3 and SLFILTERW7).

Pre-normalization data processing

For each spot, the foreground red ($R_f$) and green ($G_f$) quantitated fluorescent intensities (acquired using GenePix image analysis software) of the arrayed DNA sequences were used to compute the non-normalized log ratio, $M = \log_2(R_f / G_f)$, and average log intensity, $A = \log_2 \sqrt{R_f G_f}$. Because of the concern that local background values estimated by GenePix may add additional noise to the data [3], these values were not subtracted from their corresponding foreground values. For a given microarray, the log ratios were normalized using location-and/or scale-normalization techniques and its particular configuration (the LYMPHOMA and GASTRIC CARCINOMA studies employed microarrays with two and three distinct configurations respectively).

Normalization methods

Tables 1 and 2 summarize the 23 location normalization methods that remove none, one, or both of spatial- and intensity effect (detailed descriptions of how they adjust $M$ values can be found elsewhere [4,6,12,13,17,21,38,39]). In particular, Table 1 includes two methods that remove no spatial- or intensity-dependent dye bias: (i) NONRM neither removes any effect nor alters the distribution of $M$ values; and (ii) GMEDIAN does not remove any effect but acts as a baseline normalization method because it sets the mean or median of $M$ to zero. There are eight methods that remove either spatial- or intensity effect: (i) SLBSSCALE removes spatial effect at the PT level using lowess; (ii) SLFILTERW3 and SLFILTERW7 remove spatial effect using median filters of the block sizes $3 \times 3$ and $7 \times 7$, respectively [17]; (iii) IGBSCALE removes intensity effect at the global level; (iv) ILBSSCALE removes intensity effect at the local level and as
a byproduct removes spatial effect partially; (v) IST-
SPLINE removes intensity effect at the global level using
rank invariant set and a spline smoothing technique [38];
and (vi) QSPLINER (QSPLINEG) removes intensity
effect at the global level using spline smoothing applied to
quantiles obtained from $R$ ($G$) and using the geometric
mean of the $R$ ($G$) channels of all arrays as the target array
[13]. In Table 2, IGSGLOESS is a one-step process that
removes global intensity effect and global spatial effect,
while the remaining thirteen strategies are two-step pro-
cesses that remove both dye effect by combining methods in
Table 1.

Table 3 summarizes the three scale normalization meth-
ods used (detailed descriptions of how these methods
adjust the scale of $M$ values can be found elsewhere [4]).
WSCALE adjusts the scale of $M$ values at the PT level.
BScale adjusts the scale of $M$ values globally across all
microarrays in a data set. WBSCALE adjusts the scale
locally followed by globally, in two steps. These scale nor-
malization methods were applied to non-normalized data
(NONRM) and to data that had been normalized using
the five location methods SLLOESS, IGLOESS, ILLOESS,
ILLOESS-SLLOESS, or IGLOESS-SLFILTERW7. These
methods were selected to represent methods that remove
spatial and/or intensity effect at different levels.

**Post-normalization data processing**

For the five cancer-biology studies, examination of the
published data indicated that probes printed on different
microarrays (even those with the same configuration) were
not necessarily identical. For the $N$ microarrays associ-
ated with a given study ($N$ can be equated with the value
given for "Microarrays" in Table 4), the 41 data sets used to
estimate k-NN classifiers and to determine their
LOOCV classification errors were created as follows. Each
microarray was handled as described in "Pre-normaliza-
tion data processing" and the ensuing $M$ values were nor-
malized using the 41 distinct location and/or scale
techniques discussed above. A probe was retained for fur-
ther processing only if it was printed and present (i.e., suc-
cessfully measured and computed) in 95% of the $N$
microarrays. If a probe met these criteria, missing $M$
values were imputed using the k-NNimpute algorithm [40]
as implemented in the Bioconductor package/function
pamr/pamr.knnimpute($k=10$) [41]. Given the 41 data sets,
the $M$ values for a probe in all $N$ microarrays were
centered and rescaled to a unit norm. For LYMPHOMA,
the final dimensionality (number of probes after post-
normalization data processing) of each of the $N=81$ data
points was 6,850 ("Probes"). The 41 post-normalized
data sets for the five examined studies are available at
http://paccm.upmc.edu/BMCsup.html.

**Classification error**

Given $D$ data points, each of which is assigned to one of
$K$ categories (e.g., "normal", "DLBCL", "FL"), a LOOCV
procedure for this $K$-class data set is as follows. The data
set is partitioned into a test set of one data point and a
learning set of $D-1$ data points. The learning set is used to
train a classifier and the ensuing model is employed to
predict the class label of the test data point. This process is
repeated so that the class of each data point is predicted
using a classifier estimated from all other data points in
the data set. Classification error is the number of the
instances in which the predicted class of a data point dif-
ers from its known class. The error rate is this value
divided by the number of data points, $D$.

**k-NN classifier**

Given a $K$-class data set, the $k$-NN algorithm predicts the
class label of a test data point by first finding which of the
data points in the data set are its $k$ closest neighbors. The
classes of these $k$ nearest neighbors are examined and the
class of the test data point decided by a majority vote, with
ties being broken at random. If there are ties for the $k$th
nearest data point, all candidates are included in the vote.
Classification using $k$-NNs does not require any special
handling of multi-class data sets. A widely employed
measure of the proximity of two data points and the one
utilized here is the standardized Euclidean distance
[42,43]. Since all probes are treated with equal weight, the
classification results are affected by all the probes rather
than just a subset, as would have been the case if feature
selection had been employed. Euclidean distance has been
shown to be effective and accurate on a variety of
data sets [43,44].

The optimal number of nearest neighbors, $k^*$, was deter-
mind via leave-one-out cross-validation. An original data
set of $D$ data points was partitioned into a test set of one
data point and a learning set of $D-1$ data points. Given a
specific $k$, the $k$-NN algorithm was used to predict the class
of each data point in the learning set using the $D-2$
remaining data points. The classification error, $e_k$, of the
learning set was determined. This procedure was per-
fomed using $k=\{3,\ldots,10\}$ and $k^*$ taken to be the $k$
producing the smallest classification error, i.e., $\min_k(e_k)$. The
class of the test data point for the original data set was predicted using $k^*$ and the $k$-NN algorithm. This entire
process was repeated such that each of the data points in
the original data set was used as a test set. The classifica-
tion error of the original data was calculated. The $k$-NN
step was performed using the R class/package [45]class/
knn.cv($k=\text{"number of neighbors"}$) where "number of
neighbors" was set to $3,\ldots,10$. The prediction step was per-
fomed using class/knn($k=\text{"optimal } k\text{"}$) where "optimal
$k"$ was $k^*$.
List of abbreviations

AC: adenocarcinoma;
CCC: clear cell carcinoma;
DLBCL: diffuse large B-cell lymphoma;
FL: follicular lymphoma;
GCC: granular cell carcinoma;
k-NN: k-nearest neighbor;
LCLC: large cell lung cancer;
LOOCV: Leave-one-out cross-validation;
lowess: local regression estimation;
PC: papillary carcinoma;
PT: print-tip group;
SCC: squamous cell carcinoma;
SCLC: small cell lung cancer;
SMD: Stanford Microarray Database.

Authors’ contributions

WW designed and performed computational experiments, and drafted the manuscript. EPX participated in experimental design and in drafting the manuscript. ISM participated in experimental design and edited the manuscript. CM and MJB read and edited the manuscript. All authors contributed to, read and approved the final manuscript.

Acknowledgements

We thank three reviewers for their critical yet helpful comments. We thank Professor David Banks in the Institute of Statistics & Decision Sciences of Duke University for helpful discussion and suggestions. We also thank Aylin Rizki and Paraic A. Kenny for critical reading of the manuscript, and Kevin Peet and Pentelope Sig for proofreading the manuscript. This work was supported by the Department of Energy (OBER DEAC0376SF00098), by NIH (grant CA64786 8744), and by an Innovator Award from the Department of Defense (DAMD17-02-1-0438) to MJB. Work by I.S.M. was supported by the California Breast Cancer Research Program, National Institute on Aging, National Institute of Environmental Health Sciences, and U.S. Department of Energy under contract No. DEAC0376SF00098.

References

1. Chen Y, Dougherty ER, Bittner ML: Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images. Journal of Biomedical Optics 1997, 2:364-374.
2. Yang YH, Buckley MJ, Speed TP: Analysis of cDNA microarray images. Brief Bioinform 2001, 2:341-349.
3. Qin LX, Kerr KP: Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res 2004, 32:e571-e579.
4. Yang YH, Dudoit S, Luu P, Speed TP: Normalization for cDNA microarray data: A Semi-linear model-based analysis. J. Microarray data normalization and transformation. Nat Genet 2002, 32 Suppl 2:496-501.
5. Smyth GK, Yang YH, Speed TP: Statistical issues in microarray data analysis. In Functional Genomics: Methods and Protocols Volume 224. Edited by: Brownstein MJ and Khodursky AB. Totowa, NJ, Humana Press; 2003:111-136.
6. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display of genome-wide expression patterns. Proc Nat Acad Sci U S A 1998, 95:14863-14868.
7. Zien A, Aigner T, Zimmer R, Lengauer T: Centralization: a new method for the normalization of gene expression data. Bioinformatics 2001, 17:3235-331.
8. Cleveland WS, Devlin SJ: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association 1988, 83:596-610.
9. Dudoit S, Yang YH, Callow MJ, Speed TP: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. , University of California, Berkeley; 2000.
10. Dudoit S, Yang YH: Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In The Analysis of Gene Expression Data: Methods and Software Edited by: Parmigiani G, Garrett ES, Irizarry RA and Zeger SL. New York, Springer; 2003:73-101.
11. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S: A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 2002, 3:research0048.
12. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19:185-193.
13. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002, 30:e15.
14. Teng GC, Oh MK, Roblin L, Liao JC, Wong WH: Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Res 2001, 29:2549-2557.
15. Wilson DL, Buckley MJ, Hellwell CA, Wilson IW: New normalization methods for cDNA microarray data. Bioinformatics 2003, 19:1325-1332.
16. Kepler TB, Crosby L, Morgan KT: Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biol 2002, 3:RESEARCH0037.
17. Cui X, Kerr MK, Churchill GA: Transformations for cDNA Microarray Data. Statistical Applications in Genetics and Molecular Biology 2003, 2:Article 4.
18. Yang Y, Brass A, Hoyle DC, Hayes A, Bashien A, Oliver SG, Waddington D, Rattray M: A model-based analysis of microarray experimental error and normalisation. Nucleic Acids Res 2003, 31:e69.
19. Wang X, Heusner MJ, Wu Y, Pazi N, Ghosh S: Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. Bioinformatics 2003, 19:1341-1347.
