BIOINFORMATICS

Vol. 20 no. 1 2004, pages 105–114
DOI: 10.1093/bioinformatics/btg385

Adjustment of systematic microarray data biases

Monica Benito[1], Joel Parker[2], Quan Du[5], Junyuan Wu[2], Dong Xiang[2], Charles M. Perou[2,3,4,*] and J. S. Marron[6,*]

[1]Department of Statistics and Econometrics, University of Carlos III, Madrid, Spain, [2]Lineberger Comprehensive Cancer Center, [3]Department of Genetics and [4]Department of Pathology and Laboratory Medicine, University of North Carolina, Chapel Hill, NC 27599-7264, USA, [5]Department of Molecular Medicine, Karolinska Institutet, S 17176 Stockholm, Sweden and [6]Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

Received on April 4, 2003; revised on July 3, 2003; accepted on July 12, 2003

ABSTRACT

Motivation: Systematic differences due to experimental features of microarray experiments are present in most large microarray data sets. Many different experimental features can cause biases including different sources of RNA, different production lots of microarrays or different microarray platforms. These systematic effects present a substantial hurdle to the analysis of microarray data.

Results: We present here a new method for the identification and adjustment of systematic biases that are present within microarray data sets. Our approach is based on modern statistical discrimination methods and is shown to be very effective in removing systematic biases present in a previously published breast tumor cDNA microarray data set. The new method of 'Distance Weighted Discrimination (DWD)' is shown to be better than Support Vector Machines and Singular Value Decomposition for the adjustment of systematic microarray effects. In addition, it is shown to be of general use as a tool for the discrimination of systematic problems present in microarray data sets, including the merging of two breast tumor data sets completed on different microarray platforms.

Availability: Matlab software to perform DWD can be retrieved from https://genome.unc.edu/pubsup/dwd/

Contact: marron@email.unc.edu; cperou@med.unc.edu

Supplementary information: The complete figures that represent the cluster diagrams in Figure 6 and other figures are available at https://genome.unc.edu/pubsup/dwd/

1 INTRODUCTION

DNA microarrays are a powerful tool for the study of complex systems and are being applied to many questions in the biological sciences. In particular, the study of human tumors using patterns of gene expression have identified many expression differences that can predict important clinical properties like the propensity to relapse (van't Veer *et al.*, 2002) or the survival outlook for a patient (Sørlie *et al.*, 2001). However, a challenge of clinical sample studies is that systematic biases due to different handling procedures are often present. Microarray experiments are often performed over many months because sample collection is prospective, with most samples being assayed soon after they are collected. Additionally, samples/tumors are collected and processed at different institutions and may be assayed using different microarray print batches or platforms or using different array hybridization protocols.

These systematic biases are manifested as differences in gene expression patterns when one set of microarrays is directly compared with a second set of microarrays. When using 'supervised' statistical analyses, systematic biases show themselves as a subset of genes that tend to be more highly expressed in one set of microarrays versus another, and a concomitant subset of genes that are lower in expression in one set versus the other. These biases can typically be identified because they perfectly correlate with non-biological properties like where the samples were isolated and processed (source bias), or what print batch of microarrays the samples were tested on (batch effect bias). As can be expected, these systematic biases compromise the integrity of the data, and are especially troubling in experiments in which many samples are assayed over a long time period as these studies typically get assayed on many different print batches of microarrays.

Other researchers have used Singular Value Decompositions (SVDs) to correct for systematic biases in a data set of yeast cell cycle experiments (Alter *et al.*, 2000), and to correct for microarray batch bias in a data set containing many soft tissue tumors (Nielsen *et al.*, 2002). We present here a new method, called 'Distance Weighted Discrimination (DWD)' (Marron and Todd, 2002, http://www.optimization-online.org/DB_HTML/2002/07/513.html), which can be used to adjust microarray data sets to compensate for systematic biases. We examined our previously published breast tumor data sets (Perou *et al.*, 2000; Sørlie *et al.*, 2001) containing 107 cDNA microarray experiments and identified two distinct experimental biases. To evaluate the robustness of

*To whom correspondence should be addressed.

this new analysis technique, we applied DWD to this data set and showed a significant reduction in the source bias, and in the microarray batch bias. We also present data which suggests that this approach can be used to make adjustments for other systematic biases including across microarray platform effects, which suggests that DWD presents a new and powerful method for adjusting microarray data sets for systematic artifacts.

2 SYSTEMS AND METHODS

2.1 Hypothetical discrimination based adjustments

One way of understanding the problems with SVD/Principal Component Analysis (PCA) for removal of systematic effects is to recall that SVD/PCA seeks only to find 'directions of greatest variation'. When this goal is consistent with the systematic bias effect (meaning the systematic bias effect generates more variation than any other parts of the data, as measured by the sums of squares), then good results will be obtained using SVD/PCA. This appears to have driven the positive results reported by Alter *et al*. (2000) and Nielsen *et al*. (2002). However, when the magnitude of the systematic effect variation is similar to other components of variation, as is seen in Figure 5 (or perhaps even smaller as seen in Supplemental Figure 1), then this approach can easily fail. In these situations, where the first SVD/PCA direction is not appropriate for bias adjustment, a natural way to improve the analysis is to make full use of the systematic bias information (i.e. each case is known to belong to a particular batch, or known to be derived from a given source). Then instead of choosing directions to maximize variation in the full population (the goal of SVD/PCA), it is natural to choose directions to maximize separation of the bias. These points are illustrated using a hypothetical example of source effect, in Figure 1. This hypothetical example is only two-dimensional (2D) (i.e. only two genes are considered), to make it easy to visualize the data 'point cloud'. Note that the two subpopulations (shown as circles and pulses) are quite separate from each other, and have similar distributions (i.e. the same population shape), so that a simple translation would be able to remove any differences between the populations. The main goal of this paper is to identify effective ways of finding the direction (and magnitude) of this translation.

The direction vector of the first principal component (i.e. the SVD direction) for these hypothetical data is overlaid as the long thick black line in Figure 1. Note that this direction is clearly wrong for our goal of removing the difference between these populations. In particular, when the data are projected on to this direction vector, the subpopulations will overlap. The reason is that the PC1 direction is the 'direction of greatest variation in the data', which in this case is quite different from effective source adjustment. Also overlaid is the Fisher Linear Discrimination (FLD) direction. Note that this direction is correct for removal of the source effect. In particular, when each source is shifted in this direction, by an amount determined by the source subpopulation means, then the distributions will be indistinguishable. The reason that FLD works much better is that it exploits the source labels, which are ignored by SVD/PCA.

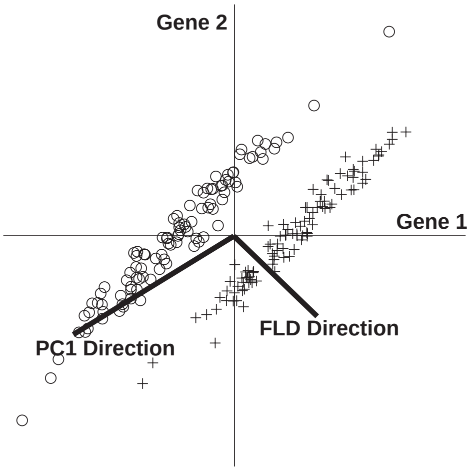

Fig. 1. Hypothetical two gene example showing how PCA directions can be wrong for source bias adjustment, thereby motivating methods based on discrimination ideas. Circles represent samples from source 1 and pluses represents samples from source 2.

In addition to finding better directions for systematic effect adjustments, we recommend another important improvement over the SVD adjustment. Instead of completely subtracting all variation in the chosen direction (as is done with the usual SVD approach), we only subtract the subpopulation means of the data projected on the given direction. This preserves any variation in this direction that is not caused by systematic effects, instead of squashing out all structure in this direction as is done by subtracting the first PC direction (particularly dangerous in SVD contexts since the first PC direction is chosen to contain 'maximal interesting structure'). In Figure 1, this corresponds to shifting the subpopulations so that they overlap, instead of projecting the data on to a single line. This is especially important when there are directions (e.g. particular genes) where there are both important bias and biological effects. Allowing overlap will keep as much as possible of the biological effects. However, there may be biological effects that are still so confounded with the bias effects that they may be diminished by this adjustment.

While FLD is very effective for the hypothetical data shown in Figure 1, it has less desirable properties for more realistic data contexts like microarrays, as is shown in Figure 2. In particular, FLD has poor performance in High Dimension, Low Sample Size (HDLSS) contexts. This problem not only arises for microarray data, but also appears in other statistical contexts, such as medical image analysis and chemometrics. HDLSS data pose a very serious challenge to

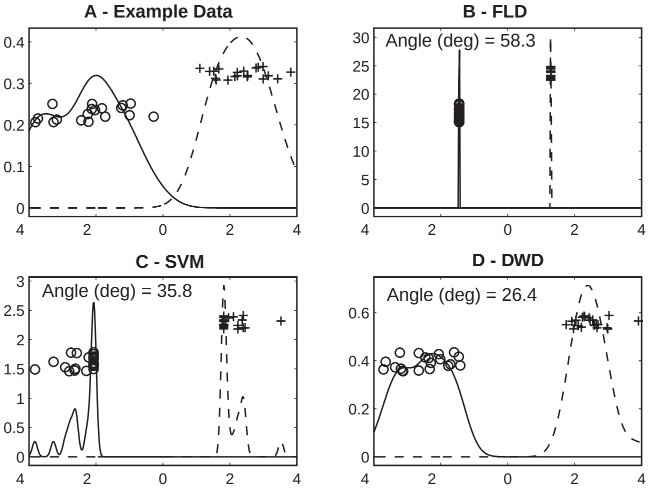

Fig. 2. (**A**) Data for 50 dimensional/gene Gaussian hypothetical example, to illustrate HDLSS failing of FLD, and superior performance of DWD over SVM. In (**B**) and (**C**) the original subpopulation shapes are lost when projecting on to the FLD and SVM directions. However, in (**D**) the Gaussian shape is retained with DWD and the angle from the optimal direction is 26.4, which is the lowest of the three. Circles represent samples from source 1 and pluses represents samples from source 2.

most classical statistical multivariate analysis settings (such as FLD), because the first step in those analyses (‘sphering the data’ by multiplying by the root inverse covariance matrix) fails, since the covariance matrix is not full rank. This point is illustrated in Figure 2A, which shows a different hypothetical example, this time in 50 dimensions/genes. The data are all simulated Gaussian, with independent components and unit variance. All the mean vectors are zero, except in the first component where there are 20 data points (shown as pluses) with mean $+2.2$, and 20 data points (shown as circles) with mean -2.2. The projections of these 50 dimensional vectors on to the first component are shown in Figure 4A, as ‘jitter plots’ [meaning random heights are used to provide visual separation of the points, Tukey and Tukey (1990)], with smooth histograms (see Wand and Jones, 1995) overlaid. While the subpopulations are clearly separated in this plot, it can be quite challenging to find this direction because of the relatively high noise level and high dimensionality (a familiar situation in microarray analysis).

Figure 2B shows the results of FLD for these data. The implementation is done with a generalized inverse of the full sample covariance matrix. The shape of the projected data sets look quite different from the projections in Figure 2A, with all the data from each class lying essentially on top of each other. This is because FLD seeks to find the direction that maximizes the separation of the classes, relative to the spread within the classes. Because there are only 40 data points in 50 dimensions, it is not surprising that this type of ‘perfect separation’ is possible. However, note that the subpopulation shapes are much different from those in Figure 2A, which represents the optimal direction for discrimination (i.e. the direction that will work the best for discriminating new data). The angle of the FLD direction (i.e. 58°), to the optimal is also shown. This shows that FLD has found a spurious direction, and is driven by sampling artifacts that will change completely for a different set of data. Essentially FLD is ‘feeling random artifacts in this particular data too strongly’, and so this direction will suffer from poor generalizability as a discrimination rule. This problem can be viewed as over fitting of the data.

Another approach to this problem is to use Support Vector Machines (SVM), discussed in detail in Section 2.3. The performance of the SVM, for the 50 dimensional hypothetical data is shown in Figure 2C. Note the projected data are no longer completely piled up, and that the angle to the optimal is substantially better, reduced to 36°. However, there is still substantial data piling at the margin (the interior points where data from both classes tend to accumulate), which is quite reminiscent of the over-fitting problem of FLD illustrated in Figure 2B. Again, there is a suggestion that FLD can also be ‘feeling too many sampling artifacts’.

Marron and Todd (2002) have addressed this problem by the development of DWD, discussed in Section 2.4 and illustrated in Figure 2D. Note that now the subpopulations appear more spread (as for the optimal projection in Fig. 2A), and the direction has a smaller angle to the optimal direction, now only 26°. Because of this strong performance in HDLSS situations, DWD is recommended for both this type of systematic artifact adjustment, and for other supervised learning (i.e. statistical discrimination) tasks for microarray data. An additional advantage of using DWD for systematic artifact adjustment is that the projected subpopulation shapes look more Gaussian, so that the subpopulation means, used in the adjustment, are more appealing as notions of ‘population center’.

2.2 Microarray production, hybridizations and initial data processing

All microarrays and samples used in this study have been previously published; the experiments used in Figures 3–6 and Supplemental Figures 1 and 2 were taken from the Stanford Microarray Database (SMD) and are described in Perou *et al.* (2000) and Sørlie *et al.* (2001), and from the Rosetta Inpharmatics web site. The remaining examples illustrate the effectiveness of DWD for across microarray platform adjustments, where the goal was to combine the Stanford cDNA microarray data set with the Agilent oligo microarray data from van’t Veer *et al.* (2002) (Rosetta Inpharmatics web site).

We first performed a number of gene filtering steps before any analyses were done. First, for all data obtained from the SMD, we filtered all genes for a signal intensity of 50 or greater in both the Cy3 and Cy5 channels and insisted

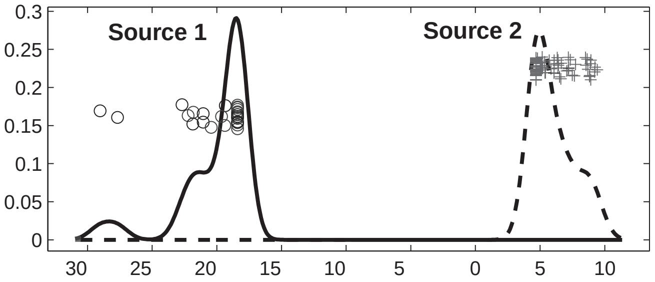

Fig. 3. Projection of data from Figure 5, on to the normal vector of the SVM separating plane shows good separation of subpopulations; however, the data are piled up at the margins.

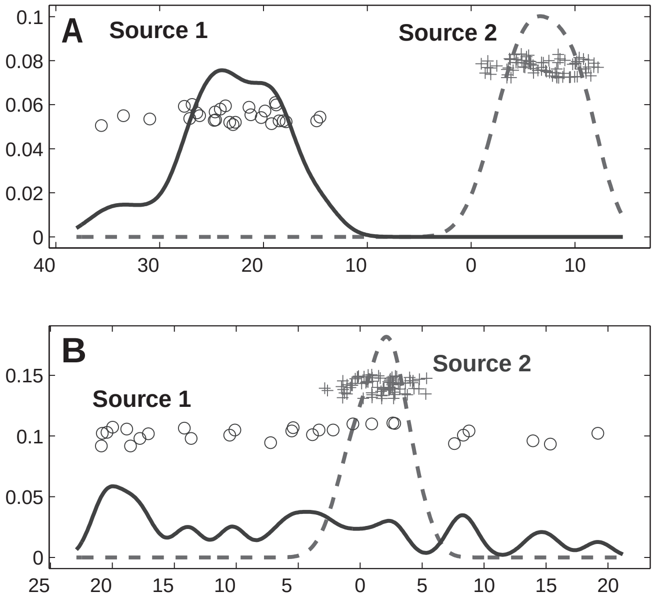

Fig. 4. Application of DWD to same data as in Figures 3 and 5. This analysis shows both good separation, and also reasonable subpopulation shape for mean shift adjustments. A = before DWD adjustment and B = after DWD adjustment.

that these signal intensity criteria be present in 70% or more of the 107 experiments for each gene. Next, we took the log 2 transformed normalized R/G ratio for each gene on the microarray. The missing values in this data table were imputed using the KNN-impute feature contained within the Significance Analysis of Microarrays plug-in (Tusher *et al*., 2001; Troyanskaya *et al*., 2001) available for use with Microsoft Excel. This imputed data set was then used for all analyses. For the van't Veer *et al*. (2002) data, we simply took their provided ratios, filtered for 70% or more of genes with provided ratios present, and imputed as described above. We linked genes from the Stanford data set to the van't Veer data set using Unigene identifiers.

2.3 Algorithms—Support Vector Machines

Support Vector Machine is a powerful discrimination method initially proposed by Vapnik (1982, 1995). Also, see Burges (1998) for an easily accessible introduction (http://citeseer.nj.nec.com/burges98tutorial.html), Cristianini and Shawe-Taylor (2000) for a detailed introduction, and http://www.kernel-machines.org/. The essential idea is to find a hyperplane that separates the two classes (i.e. each systematic bias) as well as possible. When the data are 'separable' (meaning prefect separation is possible), then the hyperplane is chosen to maximize the minimum distance of all the data to the hyperplane. The minimizing distance is called the 'margin'. An interesting view comes from studying the normal vector of the separating hyperplane, and the projection of the data upon that. This is the view shown in Figure 2C. The interior points where the data pile-up shows the margin. The SVM can be viewed as optimizing the direction vector to maximize the size of this margin. When the data are not separable, penalty terms (for those data points on the wrong side of the boundary) are added to the optimization problem, but it is still accessible to standard quadratic programming methods. The non-separable case is usually not particularly important in HDLSS situations, such as microarray analysis. This projection of the data on to the SVM normal vector, for the data of Figure 5, is shown in Figure 3. The effect is perhaps surprisingly similar to Figure 2C. Again, note that the use of the means of the projections shown in Figure 3, for adjustment in this direction, is not very attractive, because both distributions look quite skewed (in opposite directions). When means are subtracted, to adjust for the systematic effect, the population shape will be rather strange in this direction.

Note that the SVM direction represents an improvement over anything based on SVD, with the two sources far more separated than can be seen in any PC direction in Figure 5 (especially in the PC1 direction where there is considerable overlap). Thus, a major improvement of SVM over SVD for source adjustments is demonstrated for this data set. This comes from the fact that SVM is essentially aggregating over all useful directions. In Section 2.4, a further improvement, based on DWD, is proposed. This method finds a direction with a similar large spread between the batches, and gives subpopulations with a more attractive Gaussian-type shape, as suggested in Figure 2D.

2.4 Algorithms—Distance Weighted Discrimination

Distance Weighted Discrimination was initially proposed by Marron and Todd (2002). The goal is to improve the performance of the SVM in HDLSS contexts, as illustrated in Figure 2C. The main idea is to improve upon the criterion used for 'separation of classes' in the SVM. The SVM has data piling problems along the margin, because it is maximizing the

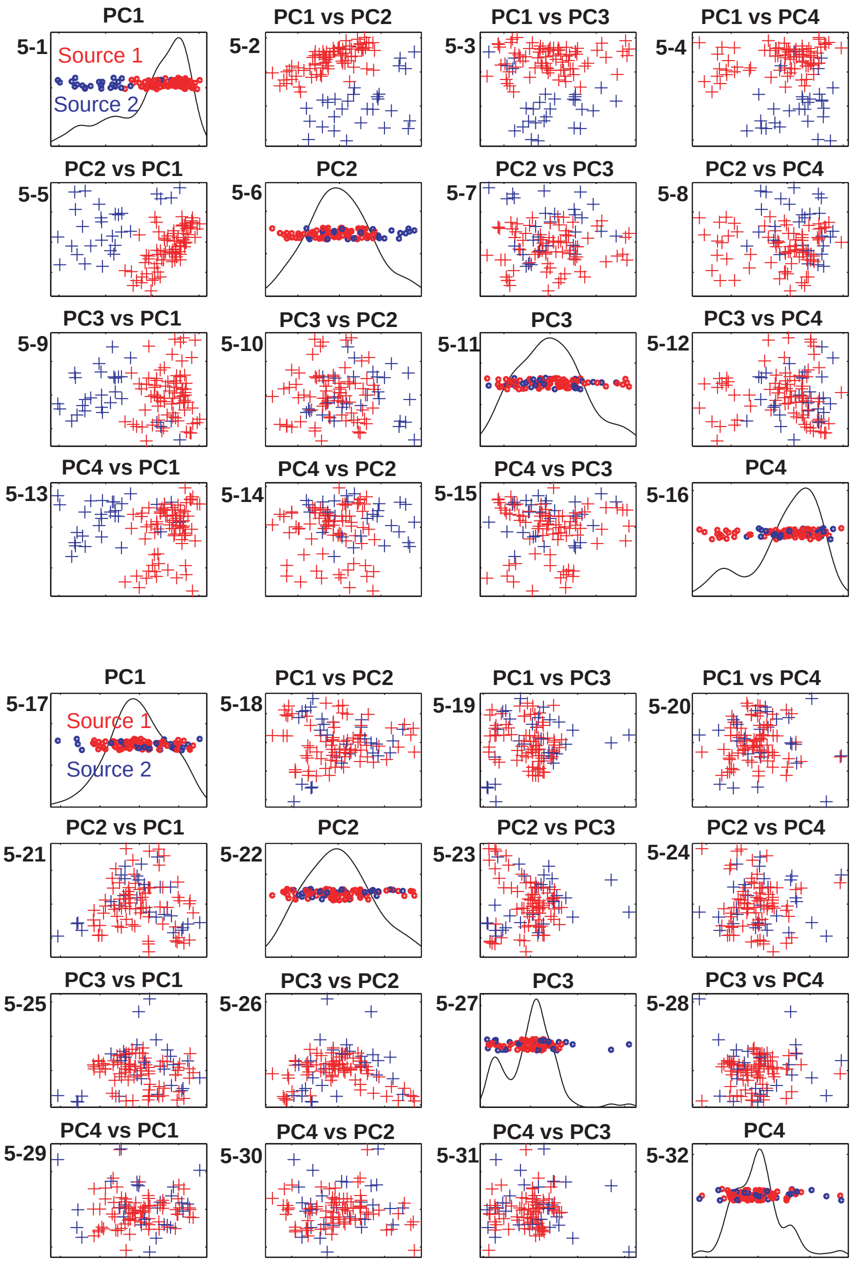

Fig. 5. 5-1 through 5-16: PCA projection scatter plot matrix of arrays, showing 1D (diagonal) and 2D projections of data on to principal component directions, of raw Stanford data. Groupings of colors indicate a source bias. The red '+' indicate samples from Norway, and the blue '+' indicate samples from Stanford University. 5-17 through 5-32: Scatter plot matrix of PCA projections, after DWD adjustment of samples from Norway and Stanford. Random dispersion of colors (instead of clustering as in the top half of Fig. 5) shows that source DWD adjustment was effective.

minimum distance to the separating plane, and there are many data points that achieve the minimum. A natural improvement is to replace the minimum distance by a criterion that allows all the data to have an influence on the result. DWD does this by maximizing the sum of the inverse distances. This results in directions that are less adversely affected by spurious sampling artifacts, as shown in Figure 2D.

Figure 4A shows the projection of the data on to the DWD direction for the same data as used in Figures 3 and 5. As one would expect from Figures 2D and 3, the sources are still well separated. A careful look at the horizontal scales shows that the 'average population separation' is even larger in Figure 4A than it is in Figure 3. Furthermore these subpopulations now look much more symmetric (even more Gaussian), so the

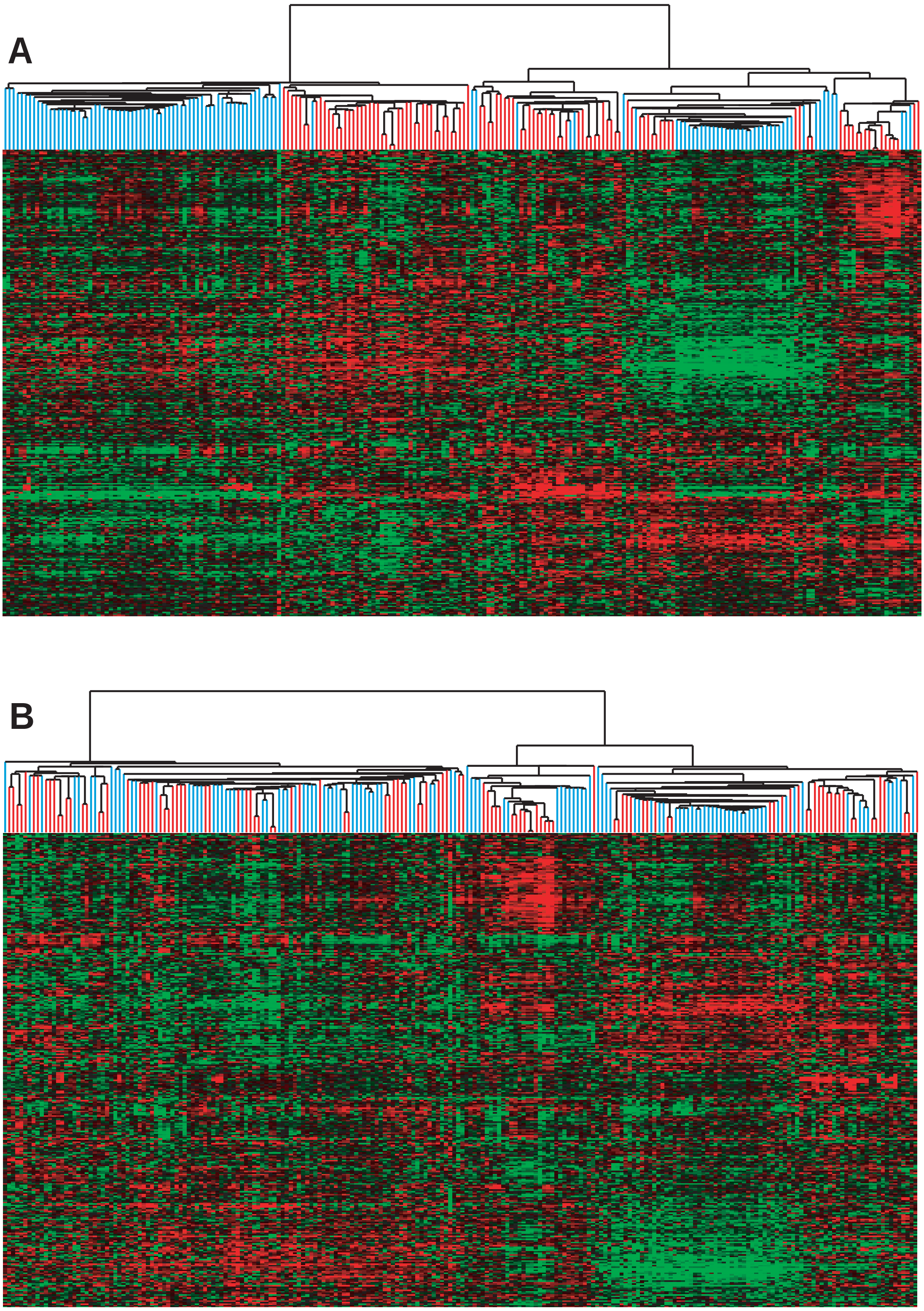

Fig. 6. Hierarchical clustering analyses of unadjusted and DWD adjusted data. The van't Veer *et al*. cases are shown in red, and the Stanford cases blue. (**A**) shows that simple median re-centering provides inadequate mixing of samples across platforms, resulting in red–green gene patterns driven in part by systematic biases. However, after the DWD adjustment resulting in (**B**), excellent inter-mixing of the cases from the different platforms/groups was seen, resulting in red/green gene expression patterns of greater biological coherence (Treeview files for before and after clusters are available as Supplemental materials).

subtraction of respective subpopulation means in this direction will remove the source effect in an appealing manner.

The specifics of the batch adjustment (thinking of the data as vectors with entries corresponding to genes) are:

(1) The DWD direction vector is found.

(2) The subpopulations (e.g. respective source subsets) are all projected in that DWD direction.

(3) The subpopulation projected means are computed.

(4) Each subpopulation is shifted in the DWD direction, by an appropriate amount, through the subtraction of the DWD direction vector multiplied by each projected mean for each gene.

Figure 4B checks the performance of DWD as a systematic bias effect removal tool, by applying the same DWD based method to the source adjusted data. Note that this time DWD does not even find a direction where the data are separated. Another verification of the good performance of DWD is the elimination of the source effect shown in Figure 6, where samples from different sources intermingle in a clustering analysis. The relative behavior of SVM and DWD shown here is very typical of a number of other examples that we have studied. Some of these are shown in Section 3 and include adjustments for microarray print batch effects, and even for microarray experiments based on different platforms.

The actual calculation of both SVM and DWD requires the use of complicated modern optimization techniques. Most implementations of SVM use quadratic programming methods, which are a set of greedy search type algorithms for solving certain convex problems. The implementation of DWD uses second-order cone methods, a broader set of algorithms, that addresses even deeper (but still convex) optimization problems. Detailed discussion of these issues may be found in Section 2 of Marron and Todd (2002), which is available with the supplemental information at https://genome.unc.edu/pubsup/dwd/

3 IMPLEMENTATION

3.1 Implementation of DWD to adjust for sample source bias

We identified in our previous microarray data set, a set of genes whose expression values very closely correlated with where the tumor samples came from (i.e. Stanford University or Norway). We do not believe that this set of genes vary due to true biological differences, but that it is instead, due to the systematic differences in how the sample RNAs were prepared. Useful views of this data can be based upon SVD, which is equivalent to PCA. Straightforward understanding of this analysis comes from thinking about the vectors of gene expressions, for each case, as points in a high-dimensional point cloud. SVD and PCA can be viewed as finding 'interesting directions' for understanding the structure of the point cloud. More precisely, they find 'directions of greatest variability'. A view that makes the 'source effect' problem clearer is shown in the top half of Figure 5. This figure shows a double matrix of plots of 1D and 2D PCA projections. The plots on the diagonal show the 1D projections (commonly called 'principal component scores') of the data on to each of the first four eigenvectors (i.e. the directions of interest in the point cloud). The individual microarray experiments are shown as colored dots, where the colors indicate the two different sources of breast tumors used in our previous studies (i.e. Norway or Stanford). The horizontal axis shows the PC scores (an axis with the numerical values is not shown because these numbers are not particularly interpretable), and the vertical axis shows a random height used for visual separation (the same 'jitter plot' visualization used in Figs 3 and 4). The black curves in the 1D diagonal projection plots are smoothed histograms (again as in Figs 3 and 4). The off diagonal graphics all show the 2D projections on to different pairs of eigenvectors (directions in the point cloud space) as scatter plots, with the x-axis corresponding to the component whose 1D projection is directly above or below, and with the y-axis corresponding to the component whose 1D projection is directly to the right or left. Thus, Figure 5-2 is a 'flip about the 45° line' of Figure 5-5, and both of these show how the first PC direction relates to the second.

Note that in Figure 5-1, the red and blue points are somewhat separated. The approach suggested by Alter *et al.* (2000) is to remove this source effect by subtracting this PC direction from the data. However, for this data set, there is substantial overlap of source effects in the PC1 direction, suggesting that deeper investigation would be useful. A stronger suggestion that this is the case comes from Figure 5-2, which compares the first and second eigen directions (i.e. PC1 and PC2). Note that better separation between the red and blue subpopulations is possible when using a diagonal separating line, rather than using a horizontal line that would be entailed from using only the PC1 direction. This casts doubt on the approach of simply removing the first principal component from the data; in particular, removal of some linear combination of the first and second directions (i.e. a slanted line in the plot) should provide a better source adjustment. This opens the question of finding other directions, which may be more appropriate for source adjustment.

A main goal of this paper is to present some improved approaches to find directions that can better separate the data than the single first PC. The result of our 'source effect' correction using DWD is shown in the bottom half of Figure 5 (Fig. 5-17 through 5-32). Now the colors, representing the two sources, are very well mixed, meaning that the systematic sample source effects in the data have been effectively removed. The same is true for higher order PC components (we have looked at orders up to 8, but these are not shown to save space). Our results are better than those where just the first eigen vector is removed, as recommended by Alter *et al*. (2000), which are summarized in Figure 5 (i.e. the plots below the top row and to the right of the first column in Fig. 5). For example, Figure 5-8 shows a strong systematic effect still present in the data. The good results in Figure 5-17 through 5-32 can be viewed as appropriately summarizing all of the directions in Figure 5-1 through 5-16 that show a need for adjustment as well as many other directions not shown here. This summarization effect is why the visual separation apparent in Figure 4 is much more than any seen in Figure 5-1 through 5-16.

3.2 Implementation of DWD to adjust for other systematic biases

In this section, additional examples are considered that show the superiority of DWD for source adjustment over SVD approaches is not a fluke of the particular data set under consideration. The first of these is another systematic microarray bias, known as the 'batch effect'. Most spotted DNA microarrays, particularly those produced at academic facilities, are physically produced in groups of 100–200 due to the number of locations that are available on the microarray robot printing platter (see the 'M guide' at http://cmgm.stanford.edu/pbrown/mguide/index.html for robot details). A given 'print run' or 'batch' of microarrays tends to show a 'batch bias', which is manifested as a set of genes whose high or low expression perfectly correlates with what print batch the sample was assayed on. This effect can be relatively small on some batches and very significant on others; however, it has been our experience that nearly every batch of microarrays shows some systematic batch bias.

Supplemental materials Figure 1 shows essentially the same PCA scatter plots as in Figure 5, using the same set of 107 breast tissue experiments, except this time the sample points are colored according to microarray 'batch' (three batches or different print runs of microarrays were used). As in Figure 5, it is clear that there is a systematic effect of batch on the structure of the data. However, note that this time, the effect appears most markedly in the fourth eigen direction (Supplemental materials Figure 1-P). It is clear that in this case the classical SVD batch adjustment (based on the first eigen direction) would be ineffective at removing this batch bias. Inspection of Supplemental materials Figure 1 may suggest replacing the PC1 adjustment by a PC4 adjustment. This solution is not ideal for two reasons. First, it requires careful human inspection and choice. Second, as noted above, much larger improvement is available from the correct aggregation over many such directions that are automatically done by DWD. An analog of Figure 4A (not shown because it is essentially the same) shows that much better separation is available in the DWD direction than in the PC4 direction.

All the methods discussed above apply to two-class discrimination, but this data set came from three different batches, i.e. three different classes. To address this additional level of complexity, which is common in many microarray data sets (e.g. samples coming from three different sources), we took a stepwise approach. An inspection of Supplemental materials Figure 1 shows that in the PC4 direction, the very small Batch 1 (red) appears more consistent with Batch 2 (green). Hence, we first made a batch adjustment between Batches 1 and 2 (combined) and Batch 3 (blue). Next, we applied the same method to the adjusted data, to separate Batch 1 from Batch 2. Because these data also have a source effect, as illustrated in Figure 1, a third step, removing that source effect as well, is also sensible. The result of the three-step process, shown in Supplemental materials Figure 2, reveals subpopulations that are now inter-mixed (i.e. the batch effect has been successfully removed). Analogs of Figures 3 and 4, for these adjustments, show quite similar lessons: the DWD gives excellent separation and good subpopulation shapes and the SVM separated similarly well. However, the SVM caused skewed projected subpopulation shapes, which are less appealing. Because the lessons are so similar to the data presented in Figures 3 and 4, these plots were not shown.

One of the most pressing challenges in the microarray field is how to combine data that comes from two different groups. In this scenario, many different systematic biases will be present including microarray batch effects (which in this case will be even greater due to different microarray platforms), source effects as each group will utilize a different source of experimental samples, different RNA extraction protocols, and other potentially unknown sources of systematic effects. As briefly discussed above, there are a number of studies that have used DNA microarrays and a two-color experimental design, to study the gene expression patterns coming from grossly dissected human breast tumors (Perou *et al*., 2000; Sørlie *et al*., 2001; van't Veer *et al*., 2002). The combined data set of Perou and Sørlie was utilized in the earlier figures and consisted of 107 samples representing 78 grossly dissected breast tumors that were assayed using mRNA with direct labeling on cDNA microarrays produced at Stanford University (and which were assayed versus a common reference consisting of a cell line pool). The van't Veer *et al*. (2002) data set contained 117 grossly dissected breast tumor samples that were labeled using the linear amplification of total RNA, and which were assayed on custom Agilent oligo DNA microarrays (and which were assayed versus a common reference consisting of a pool of 50 tumors).

Supplemental materials Figure 3 shows the PCA representation of this combined data set. Again, these two data sets are so different that simple SVD adjustment appears to offer a reasonable adjustment. However, note that both the second and third eigen directions appear to suggest some improvement (slanted lines give better separation than horizontal ones in Supplemental materials Figure 3-B and 3-C), so improvement may be possible with the DWD method over SVD. We next adjusted the data using DWD and one view of the adjusted data is shown in Supplemental materials Figure 4; note that the red and blue populations now have very good overlap, indicating a successful adjustment. Supplemental materials Figure 4 also indicates why adjustments using simple, mean based methods would not be successful; there is a substantial outlier (visible in both the PC2 and PC3 projections). The strength of DWD, over mean based methods for bias adjustment, is its reduced sensitivity to such outliers. In addition, Supplemental materials Figure 4 shows that improvements beyond those made here are possible. In particular, our method is very good at making the subpopulation 'center points' correct. However, there are other distributional aspects such as 'spread', which are not accounted for. This can be seen in several of the plots in Supplemental materials Figure 4, where the red van't Veer *et al*. population is generally 'more spread' than the blue Stanford population. Future work is occurring to address this issue.

One goal of our breast tumor studies was to identify the natural diversity of tumor subtypes present. To accomplish this goal, we identified a set of genes that we termed the 'intrinsic' gene set (Perou *et al*., 2000), which when used to group breast tumors using hierarchical clustering analysis as implemented by Eisen *et al*. (1998), identified subsets of tumors/patients that predicted overall patient survival (Sørlie *et al*., 2001). The data displays presented in Supplemental materials Figure 4 are suggestive of good integration: however, we wished to perform a combined hierarchical clustering analysis of the Stanford and van't Veer *et al*. data sets because these two data sets represent similar microarray analyses, namely two-color microarray experiments done on grossly dissected human breast tumors.

In the combined data set hierarchical clustering analysis, the common set of intrinsic genes across both data sets was determined (311 present in both data sets out of the initial 478 'intrinsic' genes); next each data set was separately imputed as described above, and then each gene was median centered within each data set. We next combined the data sets and performed a two-way average linkage hierarchical cluster analysis using the program 'Cluster' and displayed the data using 'TreeView' (http://rana.lbl.gov/EisenSoftware.htm) (Fig. 6A). The 'adjusted' and combined data set differed in that after each data set was imputed, we used DWD to adjust the Stanford to the van't Veer data set as shown in Supplementary materials Figures 3 and 4, and we then took the adjusted data and median centered each gene across all the data and clustered (Fig. 6B).

As can be seen in Figure 6A, before adjustment, there was little intermixing of the Stanford (Blue dendrogram lines) and van't Veer (Red lines) samples as judged by examination of the hierarchical cluster sample associated dendrogram (the full cluster diagrams, with complete gene names are available as TreeView files in the supplementary materials). Even when there was mixing, these samples showed low correlations with the other samples in their dendrogram branches as evidenced by the height of the branches. After DWD adjustment, however, there was a great deal more intermixing of the Stanford and van't Veer samples (Fig. 6B); in particular, the left most dendrogram branch in the unadjusted data (Fig. 6A) contained many of the estrogen receptor (ER) positive tumors and was broken into two sub-branches, each of which was almost entirely composed of samples from one source. The corresponding ER positive branch in the adjusted data (Fig. 6B) was also on the far left, and showed a much greater degree of source intermixing. In addition, the gene expression data itself showed more continuity across the luminal/ER positive expression cluster, which is the expression cluster at the bottom of Figure 6B and that contains ER itself. We have also applied this technique to merge two distinct Affymetrix breast tumor data sets together and saw similar, but less dramatic, results due to fewer systematic biases present in data sets performed on the same Affymetrix microarrays.

One potential downfall of this correction method is insufficient numbers of samples in any one group. We have found that DWD corrections work best when at least 25–30 samples are present in each group. Another potential downfall of any type of normalization or correction applied to gene expression data is that meaningful information concerning the underlying biology may be removed. The amount of information loss is difficult to quantify; however, in both corrections presented here we find that the qualitative biological structure remains. This is demonstrated through the retention of the major subtypes of breast cancer as originally defined by Sørlie *et al*. (2001), in Figure 6B as well as the retention of the gene expression patterns. These classes are distinguished by the differential expression of a small subset of genes relative to the thousands of measurements on the array. The fine structure defined by this subset remained after DWD correction of the source and batch biases (data not shown). Further, these classes were shown to occur in an independent analysis of the van't Veer *et al*. (2002) data set (Sørlie *et al*., 2003). When the platform correction is applied to these data, an additional confirmation of the subtypes of breast cancer is demonstrated with the samples independently described as basal, luminal or HER2+ are intermixed in subclusters of the same subtype despite being produced by different institutions on different platforms.

4 CONCLUSION

We have presented a new method, based on DWD, for the adjustment of various systematic differences across microarray experiment subpopulations. In many cases, the new method can provide large improvements over previously proposed methods based on subtracting the first eigen direction from the data using SVD analysis. Our new method worked well at making adjustments for a number of distinct types of systematic biases including source and batch effects. An even more powerful application: however, was the use of DWD to lessen or remove the compounded systematic biases that are present in similar data sets generated at different laboratories using different microarray platforms.

The message observed from the PCA projection visualization is that DWD successfully removed this platform effect, which was confirmed using hierarchical clustering analysis. We recommend DWD as a general approach for removing systematic bias effects from microarray data and for merging different data sets.

REFERENCES

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10106.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 955–974.

Cristianini,N. and Shawe-Taylor,J. (2000) *Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Methods Enzymol.*, **303**, 179–205.

Gollub,J. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.

Marron,J.S. and Todd,M.J. (2002) Distance Weighted Discrimination.

Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., O'Connell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. *et al*. (2002) Molecular characterization of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.

Perou,C.M., Sørlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A. *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al*. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sørlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tukey,J. and Tukey,P. (1990) Strips Displaying Empirical Distributions: Textured Dot Strips. *Bellcore Technical Memorandum*.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

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

Vapnik,V.N. (1982) *Estimation of Dependences Based on Empirical Data*. Springer Verlag, Berlin (Russian version, 1979).

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer Verlag, Berlin.

Wand,M.P. and Jones,M.C. (1995) *Kernel Smoothing*. Chapman and Hall, New York.