Visualization of Cross Platform Microarray Normalization

Xuxin Liu[1] , Joel Parker[2] , Cheng Fan[3] , Charles M. Perou[3] , J. S. Marron[1]

[1]Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
liux@email.unc.edu
marron@email.unc.edu

[2]Constella Group, Inc.
2605 Meridian Parkway
Durham, NC 27713
jparker@constellagroup.com

[3]Department of Genetics and Pathology
Lineberger Comprehensive Cancer Center
University of North Carolina
Chapel Hill, NC 27599-7264
cfan2004@gmail.com
cperou@med.unc.edu

Corresponding Author: J. S. Marron

ABSTRACT

Background: Combining different microarray data sets together, even across platforms, is considered. The larger sample sizes created in this way have the potential to generally increase statistical power. DWD has been shown to provide this improvement in some cases. But there is an apparent contradiction with the success of the DWD based approach, and earlier analyses by others claiming to show the infeasibility of across-platform adjustment.

Results: Using two NCI 60 data sets as a test bed for studying across-platform adjustment, we replicate earlier results indicating that DWD provides an effective approach to this problem, using both novel and conventional visualization methods. In addition, improved statistical power from combining data is demonstrated for a new DWD based hypothesis test. This result appears to contradict a number of earlier results, which suggested that such data combination is not possible. The contradiction is resolved by understanding the differences between viewpoints. The negative results obtained by others were based upon a gene by gene analysis, but much better insights and analyses, including understanding why DWD works, comes from the more complete and insightful multivariate viewpoint.

Conclusions: DWD is seen, using the NCI-60 cancer cell line data as a test bed, to be an effective method for cross-platform combination of microarray data. In general, multivariate data views are much more insightful and useful than only gene by gene views for understanding microarray data.

BACKGROUND

DNA microarrays have proven to be a powerful tool for many biological applications. But serious statistical challenges remain, because the data tend to be noisy. Noise in the data could be countered by running a large number of arrays, and averaging the results, but this is currently not practical because arrays costs are still relatively high. Another approach to boosting statistical power is to combine current data with previously collected data, much of which is web available.

However, as noted by Irizarry, et al. [1], hurdles to such combinations include "biases introduced during the sample preparation, manufacture of the arrays, and the processing of the arrays (labeling, hybridization, and scanning, etc.)". Even more challenging is that the data can seem especially non-comparable when they are collected using different platforms (i.e. Affymetrix versus Agilent). DWD (Distance Weighted Discrimination), developed by Marron and Todd [2], was shown to provide effective bias adjustment for all of these situations by Benito et al. [3], including effective across-platform adjustment.

Despite these promising results, there have been a number of contradictory results suggesting that these systematic biases are an insurmountable obstacle for across platform analyses (see Kuo et al. [4] , Culhane et al. [5], Parmigiani et al. [6] and Mecham [7]). The first of these is based on an unusually direct comparison of the NCI 60 Cancer Cell Line data. These data provide an excellent test bed for studying across-platform issues because gene expression of identical cell line samples was measured by both cDNA (from Synteni, Inc.; now Incyte, Inc.) and Affymetrix (Hu6800) microarrays. Details of data availability, and preprocessing that was done, are given in the Materials and Methods Section.

RESULTS AND DISCUSSION

This NCI60 test bed is used to investigate the effectiveness of DWD in the next subsection. It is seen that DWD adjustment allows combining the cDNA and Affymetrix data sets into a single homogeneous data set, which contains all the previously known biological features of these data. Multivariate projection views are quite important to both the approach, and to understanding the adjustment process.

While the visualizations strongly suggest that the DWD across-platform adjustment was successful, it does not directly settle the central issue: does the adjustment allow the combined data to have improved statistical power? This important question is addressed in the following subsection, where hypothesis tests are considered to study whether each cancer type is statistically significantly different from the rest of the data. It is seen that in almost every case, the adjusted data provide improved statistical power, thus justifying the combination of data.

Resolution of the apparent contradiction between these clearly positive results, with the negative results of Kuo et al. [4], using the same data, comes in the next subsection. There a simulated example is used that shows that important biological structure can be missed by restricting attention to only a gene by gene view. Furthermore, it is seen that a gene by gene analysis of correlation, as done by Kuo et al [4], can suggest negligible correlation between two sets of samples. Yet after DWD adjustment, the same samples can give an extremely high *multivariate correlation*, in the direction of biological interest.

An important caveat to the application of DWD is that all important biological classes need to be represented in *both* subgroups to be adjusted. Meaning, if biological type 1 lies completely in the first group, and type 2 lies completely in the second group, and DWD is applied to this data, then DWD will eliminate the differences between groups, which means that it will also eliminate the important biological differences in the process.

ANALYSIS OF THE NCI 60 DATA

Figure 1 studies the NCI 60 Cancer Cell Line data, using a view that will be employed frequently in this paper. The most important property of this view is that it is *multivariate* in nature, as opposed to more conventional *gene by gene* views, such as commonly done for example, when using conventional heat maps (i.e. hierarchical clustering diagrams). The challenge to multivariate data views is that the human perceptual system is only capable of 1, 2 or 3 dimensional views. An approach to this issue is to focus attention on 1 and 2 *dimensional projections*, that are based on *carefully chosen directions of interest*, and which are chosen from the many that are possible to consider. Principal Component Analysis (PCA) is a commonly used method for finding directions of interest. This gives a set of multivariate directions, which are orthogonal to each other, and frequently provide useful views because these are the directions that maximize the spread of the projected data. But other directions can be very useful as well, particularly to highlight known differences of various types in the data. An example of this is shown in the step by step illustration of the DWD batch and source adjustment, available from the "DWD Bias Adjustment of Batch and Source Effects" link on the Detailed Graphics web page given above. In particular, while PC directions are often useful, for some purposes it is very insightful to include DWD direction vectors as well.

In Figure 1, the chosen directions are the first 4 PC axes, where the Principal Components have been computed using the full data set. The plots on the diagonal show the 1 dimensional projections of the data. The cDNA observations appear as green plus signs (each plus sign represents one sample, i.e. array), and the Affymetrix data are purple circles. The axis shows the projections of the data on each PC direction vector. A random height is added to each symbol just for convenient visual separation (essentially the "jitter plot" idea of Tukey and Tukey [8]). Also included in each plot is a smooth histogram, colored according to the microarray platform. Note that PC 1 points essentially in the direction of the platform difference, because this is the direction of greatest variation in the combined data set. The off-diagonal plots are projections of the data onto two dimensional planes, determined by the various pairs of the PC directions. These are scatterplots, where the horizontal axis shows the 1 d projection that appears in the same column, and the vertical axis shows the 1 d projection in the same row. Again symbols correspond to samples, and the symbol type and color indicate the platform.

Also present in these scatterplots are line segments connecting the samples from common cell lines recalling that the same cell lines were assayed on each platform. Follow the Scatter Plot View of Micro-Array Data link on the web page [9]for a detailed, step by step introduction to this type of graphic. The other PC directions show other types of structure, which will be discussed in detail below.

Already apparent in Figure 1 is some suggestion of biological clusters; for example in the PC 2 vs. PC 3 scatterplot (second panel in the second row) there is a cluster that seems to separate itself from the rest. However, the cluster is not very distinct in the sense that the distances between the two platforms is as large as the separation of the cluster from the main body of data. Another potential cluster seems to appear in the PC 2 vs. PC 4 scatterplot (last panel in the second row), but again the across-platform distances are very large relative to the cluster separation from the main data.

Perfect across-platform adjustment would change the data in such a way that each cDNA sample (green plus) would coincide with its corresponding Affymetrix sample (purple circle), and each connecting black line would have a length of 0. Of course this is impossible because these measurements were made in the presence of noise. However, in all of the off diagonal panels visible in Figure 1, the black line segments do follow intriguingly simple patterns, suggesting that in fact, some relatively simple operations could yield considerable overlap of the desired type. DWD and some subsequent adjustment steps, are aimed at accomplishing this goal.

Figure 2 shows (using the same view) the same data after DWD adjustment. This adjustment used both DWD to find the right direction for shifting the data, followed by a column-wise standardization, which is important to correctly handle scale differences present across-platforms. DWD alone would be sufficient if the connecting black lines were all parallel. A careful, step by step, visual display of the steps in this adjustment is available on the web page [9]. Note that in all of the PC directions, the huge difference between the cDNA and Affymetrix data visible in Figure 1, has essentially disappeared. The black dashed line segments, which connect samples from the same cell line, show that there are both some systematic differences , in the sense that many of the nearby line segments are approximately parallel, and some pure noise differences, reflected by nearby line segments lying in much different directions. But the key observation is that both types of noise are smaller in magnitude than the distinct clusters that are visible in the data. These clusters represent important biological structure in the data (as detailed below), which shows that the DWD normalization has reduced differences in the data to a level which is less than the biological features in this data. This is the key to effective combination of data from across different statistical platforms.

The biological significance of the clusters visible in the combined data, is studied in Figure 3. This is the same as Figure 2, except that now the two clearly visible clusters are colored. There is a red colored cluster, which shows up clearly in all of the PC 3 views (3rd column and 3rd row), and a blue cluster, which shows up clearly in the PC 2 vs. PC 4 scatter plot in the last column and the second row. The names of the cDNA arrays for the data in these clusters are shown in Table 1.

Table 1: Sample Names, in the two highlighted clusters

Red Cluster	Blue Cluster
BREAST.MDAMB435	LEUK.CCRFCEM
BREAST.MDN	LEUK.K562
MELAN.MALME3M	LEUK.MOLT4
MELAN.SKMEL2	LEUK.HL60
MELAN.SKMEL5	LEUK.RPMI8266
MELAN.SKMEL28	LEUK.SR
MELAN.M14	
MELAN.UACC62	
MELAN.UACC257	

The samples in the red cluster (left hand column) in Table 1 are all melanoma cell lines except two, which was previously shown to be a very strong cluster, that is very noticeably different from the other cancer types when using hierarchical clustering analysis (Ross et al. [10]). Also note that two breast cancer cell lines also appear in this cluster, which again repeats the previous observations of Ross et al. The points in the blue cluster are all Leukemia Cell Lines that are derived from blood lymphocytes. This is a second dominant expression patterns that reflects cell type identify, and was also identified by Ross et al. [10]. Both of these clusters are further studied using conventional heat map views (see http://genome-www.stanford.edu/nci60/images/figure1.html), and see these at the “DWD Across-platform Adjustment of the NCI 60 Data” link on the Detailed Graphics web page.

A much different view of the DWD adjusted NCI 60 data, which particularly focuses on the known biological clusters, is shown in Figure 4. The scatterplots shown in Figure 3 are informative, but the directions used in the projection view are PCA directions, which are attuned to “maximal variation (in the projected data)”. This direction frequently correlates well with important biological insights, but does not do so explicitly. While the Melanoma and Leukemia clusters appear quite clearly, the other cancer types are not easy to see, even when more PCs are studied. A completely different application of the DWD direction vector (from providing the key to bias adjustment as done above), is to provide directions that more directly target biological interest, as is done in Figure 4. The directions used there were computed by grouping the 8 biological subtypes into pairs, as shown in the axis labels for each panel. For each pair, DWD was used to find direction vectors aimed at separating the two classes from each other.

Figure 4 shows that DWD was generally very successful in providing directions which drew strong distinctions between most biological classes and the remaining data. In particular, for most classes there are considerable gaps between those clusters and the main body of the data. Not surprisingly, the Melanoma (shown in green) and Leukemia (shown in cyan) clusters have the largest such gaps. Two exceptions to this are the Non Small Cell Lung Cancers (shown in black) and the Breast Cancer (shown in Red). This is likely due to the biological heterogeneity present in these groups, which for the example of the Breast Cancer cell lines, contains lines with luminal and fibroblast-like characteristics, see Ross and Perou [11].

But the main result of Figure 4 is that for all of the biological classes, the differences between platforms (shown as black connecting lines between the pairs of symbols representing each common sample) are much smaller than the biological differences between the biological classes. Thus it is not surprising that in Section 3 it will be seen that combining data produces improved statistical power.

Note that the axes shown in Figure 4 are not orthogonal to each other, unlike for the PCA based views shown above. For this reason, plots below the diagonal are also included, because they are different (while for the PCA directions the plots were just transposes, which added no new information, and hence were not included). While this visualization builds a strong case that the DWD across-platform adjustment has been successful, it still does not directly consider the question of chief concern: is there value added, in terms of statistical power, from combining these data sets using DWD? This question is answered affirmatively in the next section.

IMPROVED STATISTICAL POWER

In this section, the focus is on the statistical problem of understanding which of the biological subtypes are statistically significantly different from the rest of the data. Figure 4 suggests the Melanoma and Leukemia clusters are clearly distinct, and the Non Small Cell Lung and Breast Cancer clusters are likely not distinct. But what about the less clear cut clusters? How can these ideas be quantified in terms of p-values?

The DWD direction is used once again, in a different way here. This time, for each class the DWD direction that best separates it from the rest of the data is computed. Statistical significance is computed by projecting the data onto the direction vector, and then computing a two sample t statistic.

The lower left panel of Figure 5 shows the names of the 8 cell lines that were labeled as Renal Cancer. The name shown in gray, RENALSN12C, had an expression pattern that was much different from the others, possible due to a mislabeling of the cancer type, so it is not used in the analysis presented here. The upper left panel shows the projected data (again colored green for cDNA and purple for Affymetrix), where the DWD direction for separating the Renal data from all of the rest is used. Note that the Renal data (highlighted in red) are quite distinct from the other data. We assess statistical significance of the Renal cluster in terms of the difference in means between the Renal data, and the rest. Thus, also shown are the values of the two sample t-statistic, for the combined data (red), for the Affymetrix only data (purple), and for the cDNA only data (green). Note that the combined t-statistic is larger than the others, suggesting that the combined data have more statistical power than either individual platform. Another feature of this plot is the Affymetrix t-statistic is larger than for the cDNA suggesting more statistical power for the Affymetrix data. It is important to resist the temptation to compare this number with the usual t distribution quantiles, because the DWD direction vector has a tendency to strongly magnify this statistic. While the t-statistics contain some information about relative statistical power, this comparison is unfair to the individual platforms because the direction vector was chosen for the combined data, which could be different from the DWD direction vector for the individual samples. This issue is addressed in the center panels.

The center top panel of Figure 5 is similar to the top left panel, except this time only cDNA data is considered. This is for computation of the DWD direction, for the projection, and for the computation of the t-statistic. Note that the t-statistic is now larger than the corresponding value in the top left panel, which shows that indeed the comparison between combined and cDNA only tests is more fair when the DWD directions is recomputed. However, the combined data still appear to provide more statistical power, which again seems to confirm the value of combining data.

The center bottom panel of Figure 5 provides the same analysis, this time for the Affymetrix only data. Again there is improvement compared to using the combined data DWD direction (top left panel), but the resulting t-statistic is still inferior to that for the combined data, again suggesting the combination has been worthwhile. Again the Affymetrix only t-statistic is also larger than the cDNA only statistic, showing improved power for that platform.

While the t-statistics give some useful information, they are still not conclusive. In particular, the sample sizes for the combined data are twice the size of what they are for the individual platform data sets. Thus the t-statistics are not comparable. This problem is overcome in the right hand panels using a permutation methodology to compute actually p-values. In particular, the data are randomly relabeled, to give sub-classes of the same size as the Renal sub-population. The DWD direction is recomputed for this relabeled data and the corresponding t-statistic is computed. This process is repeated 1000 times. The values of the t-statistics are shown as dots in the right hand panels, with red dots in the top panel for the combined data, and purple dots in the bottom panel for the Affymetrix only data. The same plot for the cDNA data is not shown, since the cDNA t-statistics were always smaller, suggesting less statistical power. Note that the numerical values of the t-statistics using relabeled data are much larger (around 10 – 25) than is typical for the usual t-distribution (around -2 – 2). This is because the DWD direction seeks to strongly separate the labeled classes, so the distribution of the dots is the distribution of the t-statistic under the null hypothesis of a non-significant cluster.

In the top panel the t-statistic of 21.7, for the combined data, from the upper left panel, is compared to the null population of the red dots. Note that the actual value is larger than nearly all of the simulated dots, showing that this t-statistic is clearly statistically significant. The proportion of these could be used as an empirical P-value, but in other cases this gives the value of 0 too frequently. For better relative comparison, a Gaussian distribution is fit to the population of simulated t-statistics (the red dots), and corresponding Gaussian quantiles are used. The same method is used for computing a P-value for the Affymetrix only data in the bottom right panel. Note that the combined data P-value of 0.02 is statistically significant, while the Affymetrix P-value of 0.41 is not significant. This shows conclusively that much improved statistical power comes for distinguishing the Renal cluster comes from the DWD combination of the data, relative to testing this hypothesis on the basis of either platform alone.

Similar analyses are available for each of the other 7 cancer types among the NCI 60 data, from the “DWD Cross-platform Adjustment of the NCI-60 Data” link on the Detailed Graphics web page. A summary of the results appears in Table 2

Table 2: Summary of cluster significance results of NCI 60 Cancer types shows that combined data almost always gives improved statistical power, relative to individual platform analyses.

Type	cDNA - t	Affymetrix-t	Comb - t	Affy-P	Comb-P
Melanoma	33.1	36.5	45.6	e-8	0
Leukemia	17.6	26.0	27.7	0.007	e-7
NSCLC	15.4	23.2	22.2	0.15	0.02
Renal	15.3	19.6	21.7	0.41	0.02
CNS	13.2	17.0	18.2	0.65	0.20
Ovarian	10.8	19.1	16.4	0.22	0.26
Colon	10.4	15.6	16.1	0.80	0.51
Breast	12.5	17.9	18.0	0.54	0.21

The results from Figure 5 are summarized in the fourth row of Table 2. The other rows contain similar summaries for the other cancer types. The second, third and fourth columns summarize the projected t-statistics, with the largest for each cancer type being underlined. Note that the cDNA t-statistics are always less than the others, again suggesting lower statistical power for all of the cancer types. But for the Affymetrix vs. combined comparison, the results are very close, with the former being better for three cancer types and the latter for four types, while they are very close for the remaining type. The P-values are shown in the final column of Table 2, with combined data P-values almost always (except for Ovarian Cancer where the results are very close) being better than the single platform Affymetrix P-values. These different impressions given by the t-statistic and by the P-value show that it is important to do the more complex permutation test, in order to fully understand the improvement of the combined data over single platform data, in determining cluster significance.

Note that different cancer types yield different levels of significance. The Melanoma and Leukemia clusters, seen to be very strong in Figure 3, are clearly significant in both the combined, and the single platform tests. For two types (Non Small Cell Lung Cancer and Renal Cancer), the difference between combined and single platform data is critical, with the combined data always giving the significant result. For the other four cancer types, neither data set flags the cluster as statistically significant. As noted above, this is not surprising for breast cancer, because it is known that there are several quite different cell types present (Ross and Perou [11]).

One might object to the fact that DWD has been used for both bias adjustment, and also at the core of the hypothesis test for verification of the method. But this is not an issue because in the first step the DWD direction is subtracted out, so any potential interaction between the two will be negative.

This clear success in across-platform combination of microarray data appears to contradict the widely held view that this is impossible. In the next subsection, a toy example is presented that will resolve this apparent contradiction, by showing that it is caused by gene-by gene methods of analysis used in previous studies not providing sufficient insights into the multivariate nature of these data sets.

GENE-BY-GENE VS. MULTIVARIATE VIEWS

In this subsection, the above apparent contradictions are resolved, and it is seen that gene-by-gene analyses need to be regarded with healthy skepticism in the analysis of microarray data, because the data are intrinsically *multivariate* in nature.

The simulated data set studied here has 4000 genes (dimensions), and is intended to reflect one important biological effect, but with gene expression measured across two platforms. There are 30 samples from each platform, split evenly between the two clusters, hence 15 points in each simulated biological cluster. Each sample is generated with independent Gaussian entries (simulating gene expression values), with standard deviation one. The means of these entries is taken to be ±0.2, in such a way that there are 4 clusters, where pairs correspond to platforms, and within each pair, the clusters simulate an important biological difference. Note that the very small difference in the means of the entries is an order of magnitude less than the noise level, so that it is essentially invisible to a gene-by-gene analysis. This is seen via both a gene-by-gene scatter plot view, and by a conventional heat map, using clustering and TreeView at the “DWD Cross-platform Adjustment of the NCI-60 Data” link on the Detailed Graphics web page. However, both the simulated across-platform effect, and the simulated biological effects have been designed to be multivariate in nature, so it is seen below that these both show up, and can be adjusted for, using a proper multivariate view.

Figure 6 shows that the simulated data gives the same conclusion as that of Kuo et al. [4], when a gene-by-gene correlation analysis is done. Each green dot in Figure 6 corresponds to one dimension (i.e. gene) of the simulated data shown in Figure 7. The horizontal coordinate of the green dot is the sample correlation of the simulated expression levels of that gene, for the paired data across the platforms. The vertical coordinate is again a random value, which provides visual separation. Most of the correlations tend to be clustered around 0. There is some variation, but the amount of this is about what would be expected at random. This is essentially a reconstruction of the results of Kou et al. [4], for this simulated data.

The limitation of the gene-by-gene correlation analysis is made clear in the PCA multivariate scatterplot view of these data (shown in Figure 7). These plots show various projections on the first three principal components. Note that the first two principal components (top center panel) highlight the deliberately constructed structure in the data. In particular, the platform effect (indicated by the red and blue colors) is clear, as this strong simulated biological effect, shown as two clusters (indicated by the plus and circle symbols). The fact that platform adjustment can be successful here is indicated by the fact that black lines (connecting paired samples) are approximately parallel, which was not apparent in the gene by gene views.

After DWD adjustment, the platform effect essentially disappears (this can be seen in a plot shown on the above web page). The 4 clusters visible here become two clusters. The paired data are not exactly on top of each other, but the black connecting line segments are all much smaller than the distances between clusters. This is a simulated reconstruction of the lessons learned in Figure 4.

Finally, we revisit the correlation analysis. The limitation of the Kou, et al. [4] analysis was that it only looked in the gene by gene direction. The same suggestion of no correlation still applies for the DWD adjusted data. However, the relevant direction is not gene by gene, but instead in the PC 1 direction. When the correlation is computed on the data projected in this direction, the correlation becomes 0.98, meaning there is very strong information in this when one looks in the correct direction. This explains the above apparent paradox.

Further details of this analysis, including more detailed views and access to the data set itself, are available from the “DWD Cross-platform Adjustment of the NCI-60 Data” link on the Detailed Graphics web page. A minor technical note is that Kou, et al. [4] eliminated a few pairs of samples, because there were some questions about data quality. While that was a good decision, for the point that they were trying to make, we have chosen to include all of the data. The reason is that we wish to make the point that we *can* do across-platform normalization, and believe it is important to demonstrate this even in the presence of a few samples of questionable quality.

The ideas discussed in this section are related to the Geometric Representation ideas of Hall, Marron and Neemon [12]. That paper develops a mode of non-standard asymptotic analysis, that is relevant to High Dimension Low Sample Size data, such as microarrays. Some surprising underlying structure is shown, which is then used for classical mathematical statistical purposes, such as comparison of discrimination rules.

CONCLUSIONS

The NCI-60 Cell line data were used as a testbed to reconfirm that DWD is an effective tool for the cross-platform combination of microarray data sets. This was shown both with appropriate visualizations, and also through a careful study of improved statistical power for the combined data. The apparent contradiction with earlier published results is resolved through the fact that the previous analysis was restricted to a gene by gene analysis, while the nature of microarray data is intrinsically multivariate. This highlights the importance of truly multivariate approaches and visualizations for understanding microarray data.

MATERIALS AND METHODS

Availability: A Matlab version of DWD is available from the "Matlab software for DWD adjustment" link of the web-page [13]. A more portable version is being developed as part of the caBIG Project, [14]. Many more detailed graphics, together with explanations, are available on the web page [9].

Details of NCI-60 Data and Preprocessing: The expression values are internet available at [15] and [16]. Because the samples are identical, the effectiveness of across-platform adjustment can be precisely calibrated. Finally, because the cDNA expression values are on the scale of differences of log intensities (without the commonly used LOESS normalization in this cDNA dataset); we also work with the log, base 2, of the Affymetrix data. The original data were generated by Affymetrix Microarray Suit 4.0. There were some negative values, that were small in absolute value, which were set to 1 before taking logs. The cDNA values had some missing data points, which we handled by imputation using k-nearest neighbors imputation, see Troyanskaya, et al. [17]. We linked genes from the two NCI-60 data sets by mapping the cDNA and Affymetrix identifiers to Unigene Cluster Identifiers (UCID). Duplicate UCIDs were collapsed by taking the median value within a sample. The combined data set was created from the intersection of these two sets of UCIDs.

LIST OF ABBREVIATIONS

caBIG – Cancer Bioinformatics Grid
cDNA – Complementary Deoxyribonucleic Acid
DWD – Distance Weighted Discrimination
LOESS – Locally Weighted Scatterplot Smoother
NCI – National Cancer Institute
PC – Principal Component
PCA – Principal Component Analysis
UCID - Unigene Cluster Identifiers

ACKNOWLEDGEMENT

The research of X. L. and J. S. M. was partially supported by NSF Grant DMS-0308331, and the NCI caBIG Program. C. M. P. was supported by funds from the NCI Breast SPORE program to UNC-CH (P50-CA58223-09A1) and by the National Institute of Environmental Health Sciences (5-U19-ES11391-03).

REFERENCES

1. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.** *Biostatistics* 2003, **4:** 249-264.

2. Marron JS, Todd, MJ: **Distance Weighted Discrimination**, [http://www.optimization-online.org/DB_HTML/2002/07/513.html]: 2002.

3. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS: **Adjustment of systematic microarray data biases.** *Bioinformatics* 2004, **20:** 105-114.

4. Kuo PW, Jenssen T–K, Butte AJ, Ohno-Machado L, Kohane, IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18:** 405-412.

5. Culhane AC, Perrière G, and Higgins DG: **Across-platform comparison and visualization of gene expression data using co-inertia analysis.** [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=317282&rendertype=abstract]: 2003.

6. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer.** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15131026]: 2004.

7. Mecham BH, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane, IS, Szallasi Z: **Sequence-matched probes produce increased across-platform consistency and more reproducible biological results in microarray-based gene expression measurements.** [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=419626]: 2004.

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

9. **DWD Bias Adjustment Graphics Page** [http://genome.med.unc.edu:8080/caBIG/DWDindex.htm]

10. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JCF, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D. and Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nature Genetics* 2000, **24:** 227-235.

11. Ross DT, Perou CM: **A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines.** *Disease Markers* 2001, **17:** 99-109.

12. Hall P, Marron JS, Neeman A: **Geometric Representation of High Dimension**

Low Sample Size Data,
[http://www.stat.unc.edu/postscript/papers/marron/GeomRepn/hm5.pdf.]: 2004.

13. **Adjustment of Systematic Microarray Data Biases**
[https://genome.unc.edu/pubsup/dwd/]

14. **Cancer Biomedical Informatics Grid** [http://cabig.nci.nih.gov/]

15. **NCI Genomics and Bioinformatics Group Microarray Data Sets: cDNA,**
[http://discover.nci.nih.gov/datasetsNature2000.jsp]

16. **NCI Genomics and Bioinformatics Group Microarray Data Sets: Affymetrix,**
[http://discover.nci.nih.gov/datasetsPnas2001.jsp]

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

FIGURE LEGENDS

Figure 1: PCA scatterplot view of raw (log scale) NCI 60 data. Dashed lines connect identical samples. This shows a large difference between measurements made by the two platforms, so some adjustment is essential before combining data sets. Yet differences are systematic, which offers hope of careful adjustment.

Figure 2: PCA scatterplot view of DWD adjusted NCI 60 data. Shows effective removal of platform bias. In particular, distances between same cell lines (shown as back dashed lines) are now much smaller than distances between apparent clusters.

Figure 3: PCA scatterplot view of DWD adjusted NCI 60 data. Important biological clusters are highlighted: Melanoma (red), Leukemia (blue)

Figure 4: Scatterplot view of the DWD adjusted NCI-60 data, this time using DWD direction vectors. Different colors indicate cancer classes. This shows that across-platform differences are predominantly much smaller than differences between cancer types.

Figure 5: DWD – Permutation based hypothesis tests of statistical significance of the Renal Cancer Cluster. Main lesson is that the combined data statistical inference is more powerful than for the Affymetrix only data.

Figure 6: Gene-by-gene correlation analysis of simulated data. Shows no significant correlations, as in the earlier analysis of the NCI-60 data.

Figure 7: PCA scatterplot view of simulated data. Shows simulated strong platform and biological effects are present in these data.