

**Inter-gene correlation on oligonucleotide arrays: how much does
normalization matter?**

David L. Gold, Jing Wang and Kevin R. Coombes*

Departments of Biostatistics and Applied Mathematics, M. D. Anderson Cancer
Center, The University of Texas, 1515 Holcombe Blvd., Houston TX 77030-4009.

*Address correspondence to:

Dr. Kevin Coombes

Head of Bioinformatics Section

Department of Biostatistics, Unit 447

The University of Texas M. D. Anderson Cancer Center

1515 Holcombe Blvd.

Houston, TX 77030-4009

713-794-4154 (voice), 713-794-1915 (fax)

Key words: Microarray, Normalization, Preprocessing, Gene Correlation

Running Title: Impact of normalization to inter-gene correlation

Abstract

Normalization is a standard data preprocessing procedure in microarray data analysis to minimize the systematic technological variations in order to produce more reliable results. A variety of normalization approaches have been introduced and are widely applied. Normalization, however, remains controversial. The sensitivity of array results to normalization is an open question. No clear standard criteria for comparing or judging normalization methods has yet emerged, and the effect of normalization on gene-to-gene co-expression within an array is not clear. In this investigation, we applied several normalization methods to simulated data and real microarray datasets and evaluated the effect on gene-to-gene co-expression within an array.

We found that: (1) clear differences were seen in the distributions of gene-wise correlations between normalization methods, (2) increasing the quantity of standard quantiles in the final quantifications, i.e. more rigorous normalization, smoothed out unlikely trends in inter-gene correlation and (3) increasing the quantity of standardized quantiles did not markedly reduce the correlation of known overlapping targets. In conclusions, normalization plays a very important role in the estimation of inter-gene dependency. Caution should be had in making inferences concerning gene-wise dependencies with microarrays until this source of variation is better understood.

Availability: Oligonucleotide array data are available as CEL files from MIT's array repository: <http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi>. The DNA-Chip

Analyzer (dChip) <http://biosun1.harvard.edu/complab/dchip/>. R scripts are available upon request from krc@odin.mdacc.tmc.edu.

Introduction

The knowledge of gene expression has broad applications in basic biological sciences for understanding the cellular processes in biological systems and in clinical sciences for developing new diagnostic methodologies and treatment strategies. The development of microarray technologies has launched new expectations of ever larger-scale gene expression studies and has made a strong impact on life sciences research. Already, clinical trials are incorporating microarrays as standard practice (Liu and Karuturi 2004). In spite of these advances, however, many open questions remain in the arena of microarray analysis, including the choice of preprocessing and normalization techniques.

In order to obtain accurate gene expression measurements, random and systematic variations associated with microarray technology must be accounted for, typically by normalization. Although widely applied, normalization of microarray data remains controversial. The sensitivity of array results to normalization is an open question. No clear and standard criteria for comparing or judging normalization methods has yet emerged, although attempts have been made to understand the problem (Bolstad et al 2003). The need for array normalization is supported by the belief that differences in distributions between arrays result from technology rather than biology (Dudoit 2000). Systematic differences between arrays, if due to technology, are believed to induce data outliers and to distort gene-to-gene relative rankings and co-expression. Systematic differences between arrays, however, could also result from varying levels of transcript regulation between tissues or from RNA preparation methods (Gold et al 2004 and Stivers et al 2003).

Mathematically, normalization can be viewed as a dual-space problem. While the application of a normalization procedure transforms the p -point cloud of arrays over n -gene space, one cannot ignore the effect that this will have on the n -point cloud of genes in p -array space. For instance, any systematic bias in the distributions between arrays not accounted for and removed during normalization will surface in the gene-to-gene covariances. The effects of over-normalization might induce systematic differences, which will become evident in the gene space. The dual nature of this problem has just begun to receive attention (Bolstad et al 2003).

Basic science applications of microarray data include marker selection and hypothesis generation. These applications commonly use unsupervised clustering methods, which cluster genes based on pattern similarity. Systematic biases in gene signal, induced from external factors such as technology, increase the odds of spurious gene clusters, since unsupervised methods cannot distinguish biology from technology. Beyond these basic applications, researchers are investigating clinical applications such as prediction of survival (Beer et al 2002), disease (Bhattacharjee et al 2002 and Shipp et al 2002) and outcome (Pomeroy et al 2002 and Singh et al 2002). Accurate measurements of gene co-expression are vital to clinical use, since spurious predictors offer scant hope of reliably classifying future data sets. One cannot ignore the combined effects of gene predictor bias when building multivariate predictors of outcome.

A number of studies have evaluated multiple normalization methods using different criteria. Yang et al (2002) compared within and between-array normalization methods with two-color fluorescent glass slides. They noted the

importance of normalization and recommend use of “housekeeping” genes for between-slide normalization in order to control the bias of over-normalization on differential expression. In many applications, this recommendation will prove impractical. Schtad et al (2001) found that invariant difference selection (IDS) combined with smoothing spline normalization did a better job by smoothing away systematic difference between slides with a class of stable genes. Bolstad et al (2003) examined the effect of normalization on plots of the mean log intensity vs. the log ratio (also known as M vs. A plots) and on the bias of spike-in genes compared to concentration. The trends in M vs. A plots varied by normalization method, and they found that the quantile method led to the least distance between arrays. The bias in differential expression for spike-in genes was most reduced with non-linear and quantile normalization methods, although they noted difficulties in selecting baseline arrays.

Few studies have reported on the reproducibility of gene-wise correlation. Parmigiani et al (2004) reported positive inter-gene correlation across studies as measured by their integrative correlation coefficient. They found significant evidence of overall positive inter-gene cross-study correlation. Wang et al (2004) found strong evidence for positive within-gene correlation between platforms given the same patient cohort. Inter-platform within-gene correlation tended to be positional dependent in that genes with smaller cross-platform correlation tended to be located closer to the 5' end.

The central issue here is the effect of conventional normalization techniques on gene-to-gene co-expression. If one finds a cluster of genes with high inter-gene

correlation, should it be attributed to biology, to technology, or just to normalization? When identifying gene dependencies to explain variation in survival, what are the chances that the predictors will yield reliable results in new data, generated across laboratories and times? And what are the consequences of ignoring these effects? In this article, we use both simulated and real microarray data to study the behavior of a standard measure of gene inter-dependency, the Pearson Correlation Coefficient between genes, across some of the better known normalization methods.

Microarray Datasets and Methods

(1) **Data sets:** We used 4 publicly available cancer-related Affymetrix microarray datasets for this investigation: (1) adenoma carcinoma (Bhattacharjee et al 2001), (2) central nervous system embryonal tumor (Pomeroy et al 2002), (3) diffuse large B-cell lymphoma (Shipp et al 2002) and (4) prostate tumor (Singh et al 2002). Summaries of the datasets are provided in Table 1. Each data set was dichotomized based on either outcome, tissue, or disease. Inter-gene correlation was computed within classes rather than across classes to avoid obvious differences. We only report the distributions of the inter-gene correlations for the larger of the two classes.

(2) **Data preprocessing and normalization:** Each data set was downloaded as CEL files and quantified with: dCHIP V1.3 (Li and Wong 2001) and Affymetrix Microarray Array Suite 5.0 (Affymetrix 2000). To process the data with dCHIP, we first normalized at the probe level with the invariant set method, followed by probe set signal quantification using the PM-only model of Li and Wong (2001). We used MAS 5.0 to adjust for background and then quantified probe set signal using the Tukey Bi-weight method (Hoagun, Mosteller and Tukey 2000).

In addition we applied 1-, 2- and N -quantile normalization to the MAS 5.0 signal estimates. Here, 1- quantile normalization consisted of standardizing the 75th percentile across arrays with a multiplicative constant; 2-quantile, of standardizing the 25th and 75th percentiles with both a multiplicative constant and a linear offset; and N -quantile, following the method of Bolstad (2003). The standard quantiles for 1- and 2-quantile normalization were the medians of the 25th and 75th percentiles across arrays. The standard quantiles for N -quantile normalization were the respective quantile averages over the arrays. Both untransformed and logarithmically transformed (base 2) versions of each normalized data set were examined.

For each normalization method and data set, we examined trends in the distribution of gene-wise correlation as a function of relative signal strength rank. Two ranking methods were applied. Genes were ranked differently depending on whether or not the data were log transformed. Untransformed gene expressions were ranked in the orthogonal direction of sample standard deviation to sample average, see **Figure 1**. We call this ranking method the **orthogonal-slice**. The method was motivated by our experience of observing heterogeneity in the variance of signal expression as a function of mean signal (not shown). Log transformed gene expressions were ranked using a univariate t-statistic, called **t-stat**, testing the mean against 0, $H_0: \mu=0$, again within each of the multiple tissue or outcome factors.

For each ranking method, the genes were ordered and partitioned into W mutually exclusive windows consisting of 100 genes. The inter-gene correlation matrix R_W was computed for the genes in each window and the boxplots of the distributions of these correlation coefficients were plotted as a function of W . We

interpret trends in the distributions of R_W as a function of W as artifacts of the data processing. The Pearson correlation coefficient, r_{ij} for variables Y_i and Y_j , was computed as the covariance between Y_i and Y_j , divided by the square root of the product of variances,

$$r_{ij} = \text{Cov}(Y_i, Y_j) / \sqrt{\text{Var}(Y_i)\text{Var}(Y_j)}.$$

Results

(1) Simulation Results

We examined the effects of systematic sample bias and normalization on inter-gene correlations with an uncorrelated simulated dataset. Let y_{ij} represent gene expression, where $i = 1, \dots, 1000$ denotes genes, and $j = 1, \dots, 100$ indicates samples. We simulated gene expression from the model

$$y_{ij} \sim N(\lambda, \lambda/2), \quad \lambda \sim \text{Exp}(1/10),$$

where y_{ij} follows a normal distribution with mean λ and standard deviation $\lambda/2$, and λ follows an exponential distribution with mean 10. **Figure 1** shows the trend in sample standard deviations versus the sample averages. Superimposed are several curves orthogonal to the trend to illustrate the **orthogonal-slice** ranking procedure. Each sample j was perturbed by a small multiplicative random variable β following a Gamma distribution,

$$\beta \sim \text{Gamma}(\text{shape}=1, \text{rate}=1/2)$$

$$\alpha \sim N(0, 0.1)$$

followed by a small linear Gaussian random offset α with expectation 0. These small perturbations were designed to simulate systematic differences between the samples.

Figure 2(a–c) shows the boxplot distributions of the correlation coefficients as a

function of the window W , indexed from lowest to highest ranking using the **orthogonal-slice**: (a) before perturbation, and after perturbation and application of (b) 1-quantile normalization and (c) a linear median correction. The linear median correction was applied to show the effect of simply standardized the medians across samples linearly in the presence of systematic scale differences.

The box plot whiskers were positioned to contain 95% of R . The corresponding bands were positioned at the respective 75th and 95th percentiles of a random variable following the distribution $\text{Beta}(N/2-1, N/2-1)/2$, i.e. the null Beta distribution assuming that the true correlation is 0 (Fisher 1915 and 1928). Hence, if there is no true correlation different from 0, the boxplot edges and whiskers should approximately line up with the superimposed bands.

These figures illustrate the very obvious trends possible in the distributions of inter-gene correlation as a function of window order with incomplete normalization. The boxplots for the unperturbed data show a very well behaved pattern in the correlation coefficients lining up well with the expected Beta quantiles. Positive correlation was induced by perturbations, Figure 1(b)-(c), which was not completely accounted for with normalization. A linear median correction tends to pull the distribution of correlation coefficients for genes with signal near the median signal in line with the null. Correlations for genes with signal in the extreme are left distorted. Hence, excess of positive correlation, it was postulated, would be evidence of unaccounted for random array effects rather than biological gene-wise co-regulation.

(2) Microarray Results

Microarray results are more difficult to interpret, since true inter-gene co-expression will of course be confounded with systematic differences related to normalization. This issue cannot be avoided with real data. Therefore, we mainly concerned ourselves with the identification of obvious trends between normalization methods across data sets. We return to this vexing issue below, examining the distribution of correlation coefficients between overlapping targets.

Figure 3 shows boxplots after applying dChip normalization and PM-only model based quantification. **Figure 3(a)**, for the Bhattacharjee dataset, shows on average correlation centered about zero, although the quantiles extend well beyond those expected with the bull Beta distribution. The dispersion is more extreme with the log transformed Singh data ($N = 51$; **Figure 3(b)**).

Figure 4 contains boxplots of the distribution of inter-gene correlations for the log transformed expressions of the Pomeroy dataset quantified by MAS 5.0: (a) without additional normalization, (b) 1-quantile, (c) 2-quantile, and (d) N -quantile normalization. The data quantified by MAS 5.0 alone shows a trend in positive inter-gene correlation for low signal genes. This was improved with 1-quantile normalization, but there were still signs of strong positive correlation for genes with the highest relative signal. These trends were not nearly as apparent with 2- or N -quantile normalization. Both the boxplots and the bands line up well after 2- or N -quantile normalization, indicating mild if any inter-gene correlation different from zero. Similar results were seen in the other data sets.

For three data sets (Shipp, Pomeroy, and Singh), the loess method described in Bolstad et al (2003) was also applied to the MAS 5.0 quantified signal. Figure 5

shows the boxplots for the Shipp dataset normalized with loess. These results were similar to N -quantile normalization.

These results leave open the question of whether increasing the degree of normalization (e.g., the number of quantiles standardized) reduces artifactual correlation or instead distorts the biology. To address this question, we examined the effect of normalization on the Pearson Correlation Coefficients between known pairs of overlapping targets. Figure 6 displays correlation coefficients between probesets for the Bhattacharjee data set after 1-, 2-, or N -quantile normalization. The top row (**Figure 6(a-c)**) shows the distributions of the correlation coefficients between 1,000 randomly chosen pairs of probe sets on the U95A array. The distributions in this case line up reasonably well with the null Beta distribution, tending to show slightly more mass in the tails than expected. The bottom row (**Figure 6(d-f)**) shows the distribution of correlation coefficients between probe sets with the same Unigene ID (build 170). In this case, the heavy mass in the right tail is not visually affected. It appears that increasing the number of quantiles did not distort the correlation between probe set sequences derived from the same gene.

Discussion

Normalization of microarray gene expression data plays a crucial role in downstream data analysis, including identifying expression patterns for classification and prediction. Various normalization approaches have been introduced, but the appropriateness of each method is debatable. The distributions of gene expression patterns within samples and across samples are interconnected in complex ways.

Incorrectly modifying the within-sample distributions by under- or over-normalizing can potentially distort the expression profiles across samples and bias gene-wise correlation estimates, distorting conclusions concerning gene-to-gene dependencies.

In this paper, we have looked at the effect of different normalization methods on gene-gene correlations across samples. We saw clear differences in the results by normalization method, on both simulated and real microarray data. Although biological validation of these results is non-trivial, it is clear that normalizations methods differ to the degree that they may induce or dampen the gene-wise correlations as a function of relative signal. Overall, dChip appeared to show the most dispersion in inter-gene correlation, well outside of the expected quantiles of the null Beta distribution. If one believes that few genes actually have correlation different from 0, then the results using MAS 5.0 with 2- or N -quantile normalization offer the best results. For 3 data sets (Shipp, Pomeroy and Singh), the loess normalization described in Bolstad et al (2003) produced results similar to N -quantile normalization. Since loess normalization is time intensive, it is not preferred over 2- or N -quantile normalization.

We observed that increasing the number of standardized quantiles with MAS 5.0 tended to reduce trends in the distributions of random correlation coefficients, lining these up more closely with the expected null Beta distribution. It may be argued that increasing the number of standardized quantiles across arrays leads to a dampening of true correlation. Here though, the results show otherwise; in fact, the difference in the distributions of correlation coefficients between pairs of known overlapping targets was trivial when compared to the trends seen by signal rank order.

We concluded that it is unlikely that the true biology was distorted, since it is unlikely that the genes with the weakest signals and the most variability really should show the most positive interdependencies. Based on the results reported here, we favor quantile or loess normalization methods that line up the within-sample distributions since: (1) these methods reduced gene-wise correlation trends with signal and (2) did not markedly reduce the correlation of known overlapping targets.

In summary, we found that: (1) clear differences were seen in the distributions of gene dependencies between normalization methods, (2) increasing the number of standardized quantiles in the final quantifications reduced trends in correlation by signal intensity and (3) increasing the number of standardized quantiles did not markedly reduce the correlation of known overlapping targets. These findings are one more step in the direction to understanding the affect of normalization on inter-gene dependency.

Acknowledgements

Thank you to Kenneth Hess for his insightful and helpful comments.

Figure Legends

Figure 1 Sample Standard Deviation versus Sample Average over 1,000 simulated genes. Lines superimposed orthogonal to the trend illustrate the rank partitioning of orthogonal-slice.

Figure 2 Boxplots of the Distribution of Inter-gene Correlation as a function of W in (a) Unperturbed Simulated Data (b) Perturbed Simulated Data with 1-quantile normalization (c) Perturbed Simulated Data with an additive median correction.

Figure 3 Boxplots of the Distribution of Inter-gene Correlation as a function of W in (a) the Bhattacharjee data after dCHIP processing. Gene ranking by orthogonal-slice. (b) the log transformed Singh data after dCHIP processing. Genes are ranked by **t-stat**.

Figure 4 Boxplots of the Distribution of Inter-gene Correlation as a function of W in (a) the log transformed Pomeroy data after MAS 5.0 Gene ranking by **t-stat**. (b) the log transformed Pomeroy data MAS 5.0 signal with 1-q normalization. Gene ranking by **t-stat**. (c) the log transformed Pomeroy data MAS 5.0 signal with 2-q normalization. Gene ranking by **t-stat**. (d) the log transformed Pomeroy data MAS 5.0 signal with N -q normalization. Genes are ranked by **t-stat**.

Figure 5 Boxplots of the Distribution of Inter-gene Correlation as a function of W in the Shipp MAS 5.0 signal with loess normalization. Genes are ranked by **orthogonal-slice**.

Figure 6 Histogram of inter-gene correlations between 1000 randomly chosen pairs in the log transformed Bhattacharjee data set **(a)** MAS 5.0 1-quantile, **(b)** 2-quantile, **(c)** N-quantile. Inter-gene correlations between 2,922 pairs of overlapping targets with the same Unigene (Build 170) ID for **(d)** MAS 5.0 1-quantile, **(e)** MAS 5.0 2-quantile, **(f)** MAS 5.0 N-quantile.

References

Affymetrix Microarray Suite User's Guide Version 5.0. 2000. [online]

<http://www.affymetrix.com/index.affx>.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. 2002. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med.* Aug;8(8): 816-24.

Bhattacharjee A, Richards WG, Staunton J, Li C., Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, November 20, vol. 98 no. 24.

Bolstad BM, Irizarry RA, Strand M and Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* V 19(2), 22 January, pp. 185-193.

Dudoit S, Yang YH, Callows MJ, Speed T. 2000. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Department of Statistics, University of California at Berkeley, Technical report # 578.

Fisher RA. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10 507521.

Fisher RA. 1928. The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc. London Ser. A* 121 654673.

Gold D, Coombes K, Medhane D, Ramaswamy A, Li Z, Strong L, Koo JS and Kapoor M. 2004. A comparative analysis of data generated using two different target preparation methods for hybridization to high-density oligonucleotide microarrays, *BMC Genomics*, 5:2

Hoagun DC, Mosteller F, Tukey JW. 2000. Understanding Robust and Exploratory Data Analysis, Wiley Series in probability and mathematical statistics.

Li C and Wong WH. 2001. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, *Genome Biology* 2(8): research 0032.1-0032.11

Liu E.T. and Karuturi K.R. 2004. Microarrays and Clinical Investigations, *New England Journal of Medicine* 350; 16 April 15

Parmigiani G., Garrett-Mayer E.S., Anbazhagan R. and Gabrielson E. 2004. A Cross-Study Comparison of Gene Expression Studies for the Molecular Classification of Lung Cancer, *Clinical Cancer Research* Vol. 10, 2922-2927, May 1

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES, Golub TR. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* V. 415m 24.

Schadt E, Li C, Ellis B and Wong W. 2001. Feature Selection and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data, *Journal of Cellular Biochemistry Supplement* 37:120-125

Shipp M, Ross K, Tamayo P, Weng A, Kutok J, Aguiar R, Gaasenbeek M, Angelo M, Reich M, Pinkus G, Ray T, Koval M, Last K, Norton A, Lister T, Mesirov J, Neuberg D, Lander E, Aster J and Golub T. 2002. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature Medicine*, V. 8(1).

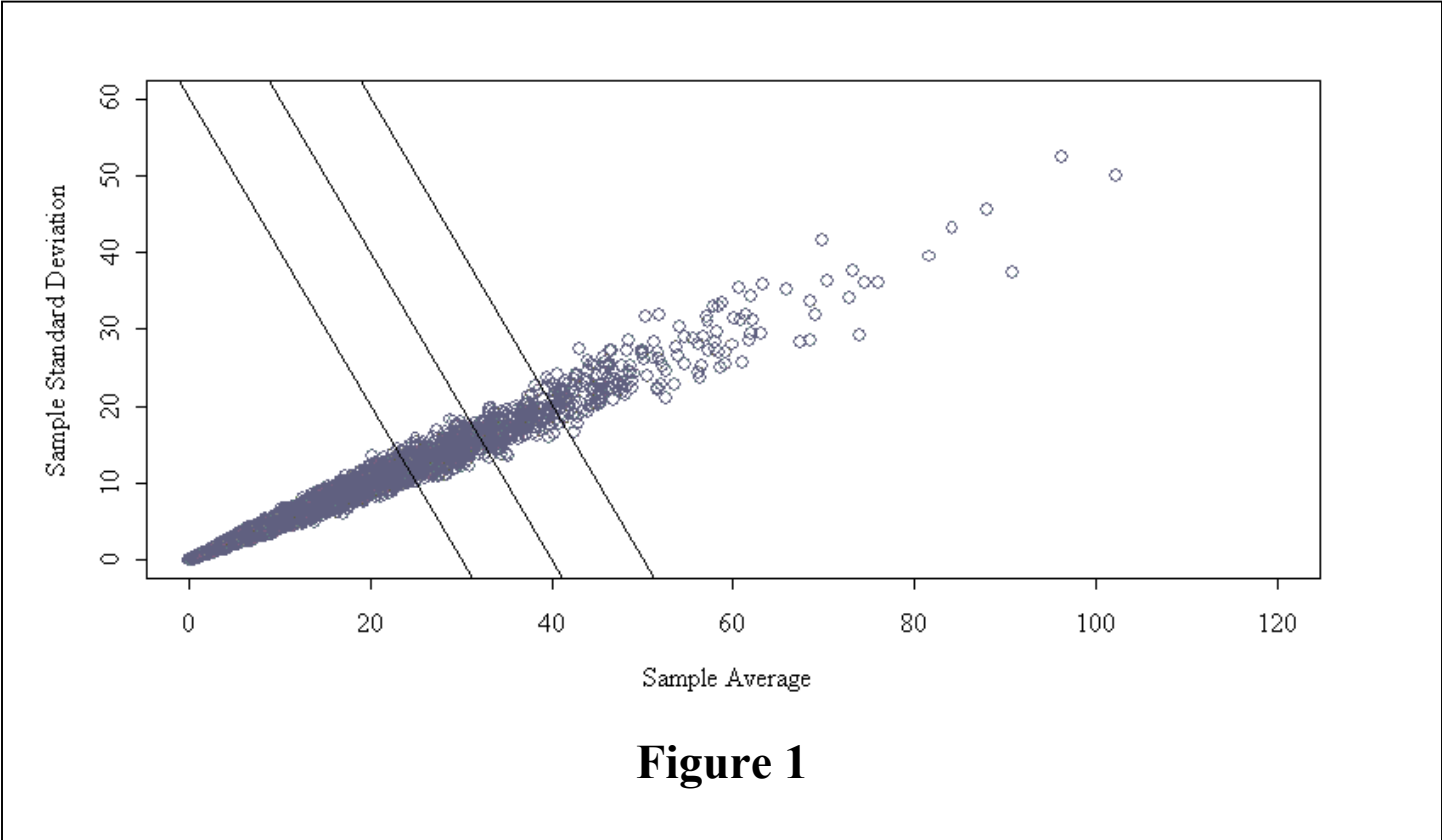
Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP, Lander ES, Loda M, Kantoff PW, Golub TR, Sellers

WR. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*: Mar. V. 1.

Stivers DN, Wang J, Rosner GL, Coombes KR. 2003. Organ-specific differences in gene expression and UniGene annotations describing source material. *Methods of Microarray Data Analysis III*, Kluwer Academic Publishers, Boston, pages 59-72.

Wang J, Stec J, Coombes KR, Ayers M, Hoersch S, Gold D, Ross JS, Hess KR, Tirrell S, Linette G, Hortobagyi GN, Symmans WF, Pusztai L. 2004. Cross platform comparison of multigene predictors of response to neoadjuvant paclitaxel/FAC chemotherapy in breast cancer generated by cDNA arrays and Affymetrix GeneChips from the same RNA, *submitted to Clinical Cancer Research*.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, Vol. 30, No. 4, e15.



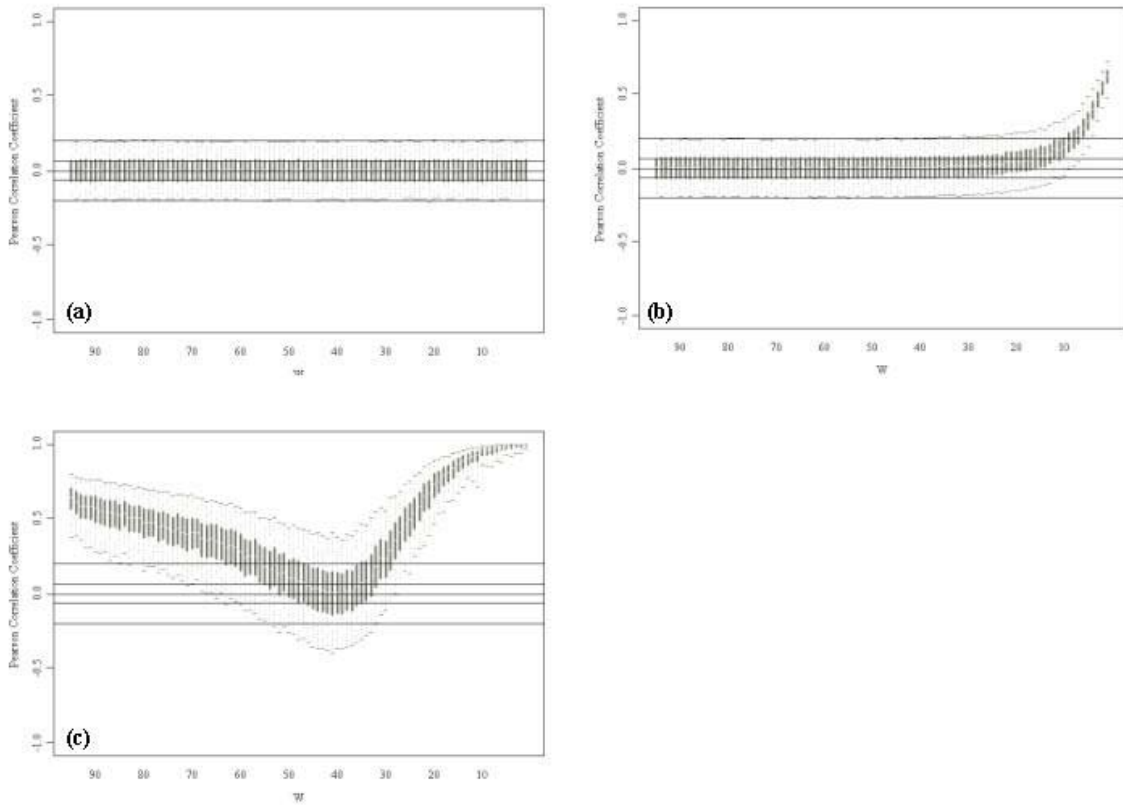


Figure 2

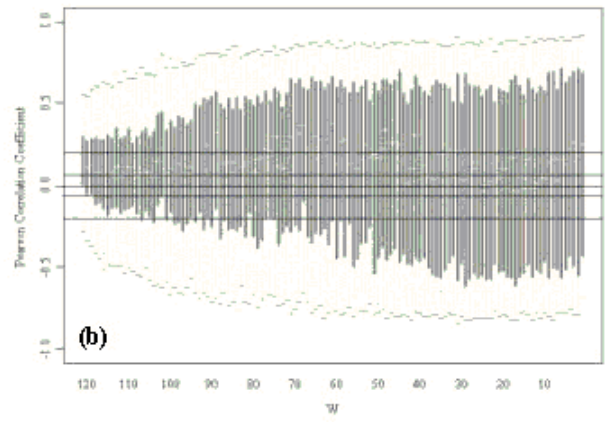
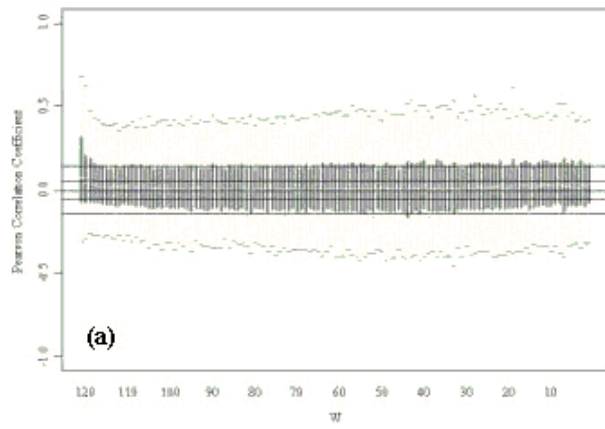


Figure 3

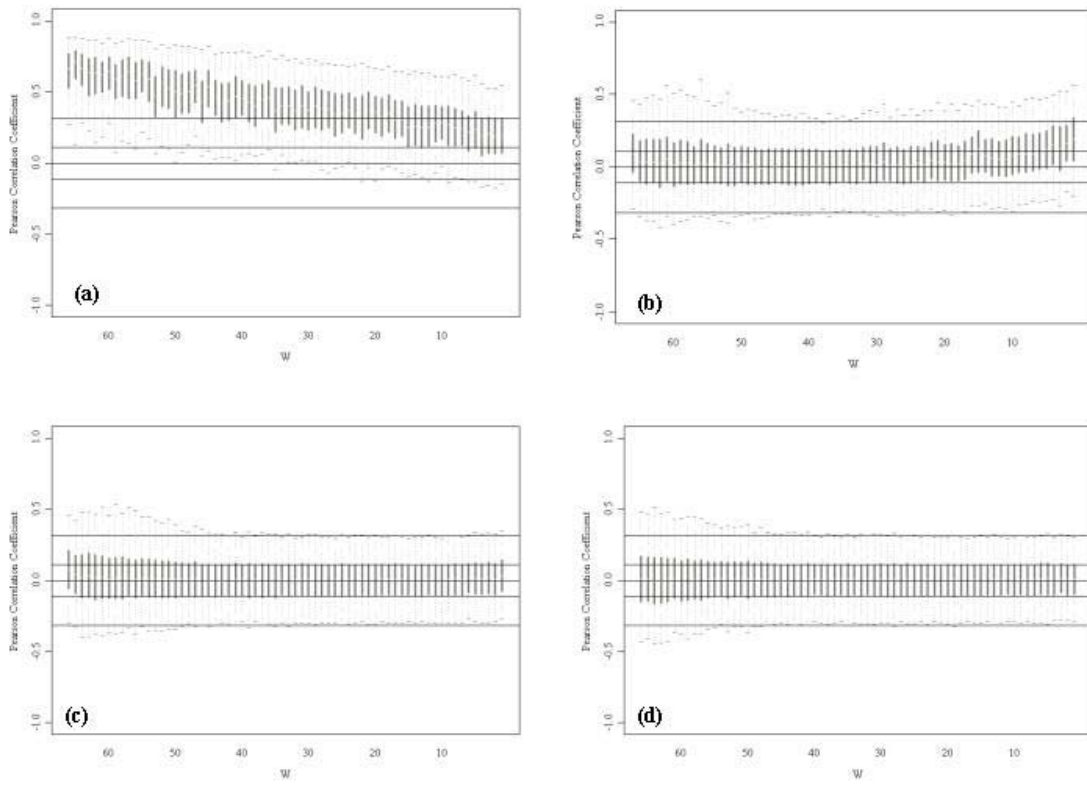


Figure 4

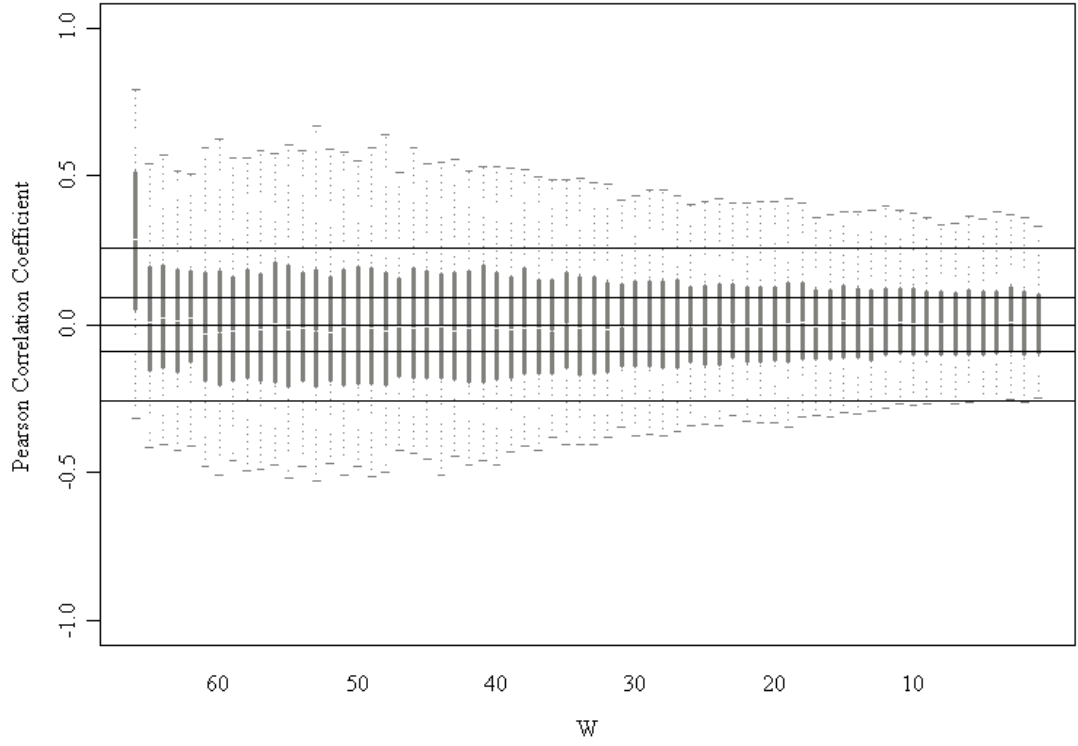


Figure 5

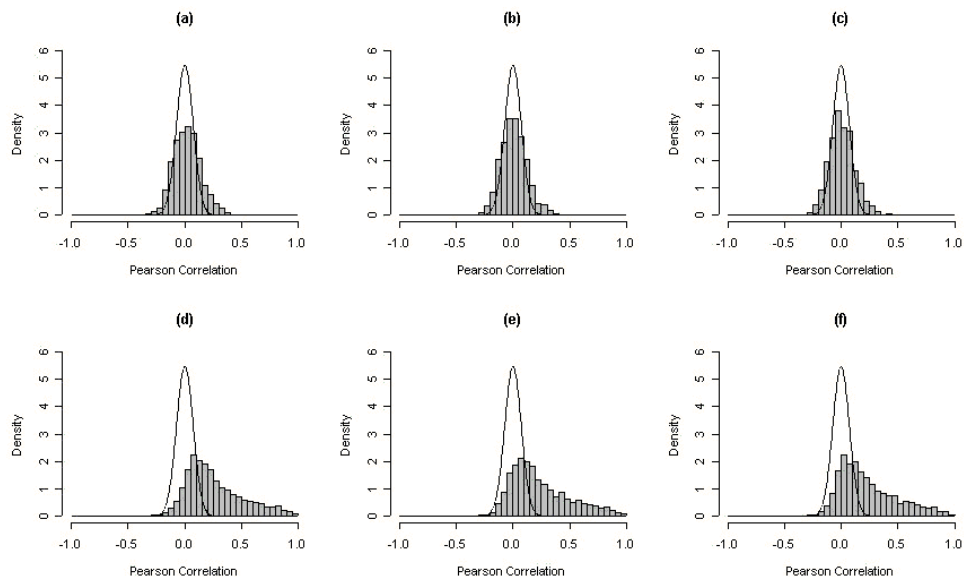


Figure 6