chip artifact *CORRECTION* (caCORRECT): A Bioinformatics System for Quality Assurance of Genomics and Proteomics Array Data

TODD H. STOKES,¹ RICHARD A. MOFFITT,² JOHN H. PHAN,² and MAY D. WANG³

¹Department of Electrical and Computer Engineering, Georgia Institute of Technology, 313 Ferst Drive, Atlanta, GA 30332-0535, USA; ²Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA 30332-0535, USA; and ³Department of Biomedical Engineering, Department of Electrical and Computer Engineering, and Winship Cancer Institute, Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA 30332-0535, USA

(Received 2 October 2006; accepted 3 April 2007; published online 26 April 2007)

Abstract—Quality assurance of high throughput “-omics” data is a major concern for biomedical discovery and translational medicine, and is considered a top priority in bioinformatics and systems biology. Here, we report a web-based bioinformatics tool called caCORRECT for chip artifact detection, analysis, and *CORRECTION*, which removes systematic artifactual noises that are commonly observed in microarray gene expression data. Despite the development of major databases such as GEO arrayExpress, caArray, and the SMD to manage and distribute microarray data to the public, reproducibility has been questioned in many cases, including high-profile papers and datasets. Based on both archived and synthetic data, we have designed the caCORRECT to have several advanced features: (1) to uncover significant, correctable artifacts that affect reproducibility of experiments; (2) to improve the integrity and quality of public archives by removing artifacts; (3) to provide a universal quality score to aid users in their selection of suitable microarray data; and (4) to improve the true-positive rate of biomarker selection verified by test data. These features are expected to improve the reproducibility of Microarray study. caCORRECT is freely available at: <http://caCORRECT.bme.gatech.edu>.

Keywords—High throughput data analysis, Chip artifact correction, Quality assurance, Biomarker, Normalization, Heatmap visualization, Bioinformatics, Systems biology.

INTRODUCTION

For microarrays to reach their full potential as a clinical molecular profiling tool for personalized and predictive medicine, the quality of microarray data

must be addressed. The FDA has started the Micro-Array Quality Control (MAQC) consortium and is seeking to develop FDA guidelines on MAQC and data analysis.^{16,22} However, the current status of MAQC and noise reduction is still a collection of scattered tools and methods. While tools such as dChip,¹⁷ RMAExpress,¹² Harshlighting,²³ and SmudgeMiner²¹ include methods to improve the quality of microarray data, these tools fail in several important aspects: (1) they do not provide sufficient visualization to help a novice user understand the source of data problems; (2) they do not incorporate spatial information into the outlier detection methods; (3) they do not incorporate outlier information into their normalization routines; and (4) they do not generate dataset quality metrics to help users select high-quality data.^{4,8}

caCORRECT (chip artifact *CORRECTION*) addresses these deficiencies and seeks to replace or complement the existing technology by combining a novel methodology with an interface that focuses on usability. caCORRECT is a web-based bioinformatics tool that uses the latest methods of normalization and multi-chip variance calculations based on technical or biological replicates to generate crisp representations of artifacts. caCORRECT then uses knowledge-based image processing techniques to flag problematic probes for replacement based on images generated using variance thresholds. Chip artifact detection is the process that the user takes to interpret heatmap visualizations and quality scores. Analysis and correction is the data processing step that takes place to replace suspicious data. The web-based design will allow public microarray archives to process large amounts of data in a short time.

To improve the quality of genomic data, it is important to understand the source of the errors and

Address correspondence to May D. Wang, Department of Biomedical Engineering, Department of Electrical and Computer Engineering, and Winship Cancer Institute, Georgia Institute of Technology and Emory University, 313 Ferst Drive, Atlanta, GA 30332-0535, USA. Electronic mail: maywang@bme.gatech.edu

Todd H. Stokes and Richard A. Moffitt contributed equally to this work.

the current state-of-art in quality control (QC). Recent studies have shown that the choice of microarray platforms is important, but not always the primary factor influencing data quality produced by laboratories. Instead, laboratory techniques are often responsible for the lack of reproducibility in microarray datasets.¹³ It has even been suggested that some gene co-expression in microarray chips is the result of spatial artifacts—with the gene pair correlations being more a function of relative chip distance than chromosomal distance.¹⁵ Even worse, the methods that are designed to alleviate such chip to chip non-uniformity could actually hamper results.^{6,9} Using caCORRECT's capability for interactive visualization, we discover several classes of artifacts which can be easily linked to their root causes. Among the most common artifacts are scratches, edge effects, and bubble effects that manifest as visible localized variations in the microarray dataset. These localized variations are not detected at the level of gene expression, but can be seen using low-level scanner outputs and by preserving the original spatial orientation of the microarray.

Much work has already been done at the gene expression analysis level to detect outlier data points and to improve the reproducibility of microarray results. Affymetrix microarrays, for example, can be processed with Affymetrix's own Microarray Suite (MAS5.0), GeneChip Operating Software (GCOS), or Probe Logarithmic Error Intensity Estimate (PLIER), but alternatives such as dChip,¹⁷ RMAExpress,¹² or GCRMA implementations in Matlab or the R statistical language (www.bioconductor.org) also exist. These programs include good measures such as normalization, background correction, and robust model fitting in an attempt to determine gene expression from multiple probe values. Many of them provide a visualization feature showing where outlier probes, or probe sets are located on the chips, but yet they do not include this spatial information in their outlier detection schemes. Direct comparisons of caCORRECT to these methods are difficult because caCORRECT is a quality assurance step happening before expression analysis, and with each of the methods mentioned above, noise removal and gene expression calculation are inseparable.

One method by Reimers and Weinstein²¹ does take spatial effects into account. This system can be used to visualize regional biases across high-density chips. Citing factors such as temperature, liquid flow rate, RNA diffusion rate, and edge effect, they showed that significant regional biases are common and can greatly affect downstream results. In addition to localized background calculation, Reimers and Weinstein's program produces a comprehensive quality score for each chip by measuring the correlation of each probe's

expression level to that of its neighbors. While this application may provide quality score information, it does not allow correction of these artifacts. Users are then faced with a difficult choice to abandon a chip, or to proceed knowing that artifacts exist.

In the development of new methods for quality control and assessment, Brodsky *et al.* proposed a novel method of using clustering of gene expression profiles across microarrays to indicate quality.⁵ First, gene expression profiles are clustered, and then the uniformity of the clusters' distributions across the microarrays are measured. Second, the patterns of high and low expressed genes are monitored on each sample for uniformity. These two methods provide a dual description of a gene's artifactual nature, which is then used to discard it from further analysis. Most of the genes discarded in this way are identified as a result of artifacts localized to one or few chips. Brodsky's strategy of removing such a gene from the experiment entirely is too harsh, and therefore caCORRECT uses a more conservative strategy to remove offending gene data from the artifact-containing chips. In this way, caCORRECT is able to retain all genes on the array while removing potentially distracting noise.

A brief survey of current microarray databases (such as GEO,²⁶ arrayExpress,²⁰ caArray, and the SMD¹) reveals that quality analysis at this level of detail is beyond the scope of many labs which produce microarrays today. The post-processing done by labs corresponds to the goals of their experiment. Some labs produce technical replicates in hopes to increase signal-to-noise ratios and to reduce the need for laborious artifact detection, while others just ignore these effects and look only at large-scale data features. For data curators and data consumers (who often only have access to the published expression data, and not to the more detailed output of the scanner), a different approach is needed. Our goal in designing caCORRECT is to make public data a knowledge resource for the whole community, and let researchers use data at a level of detail appropriate to their investigations.

METHODOLOGY

System Overview

As shown in Fig. 1, the caCORRECT workflow centers on detection and removal of localized chip variances, referred to hereafter as artifacts. The defining feature of an artifact is that it results from errors in microarray manufacturing or lab processing, not from underlying biological phenomena. The first step in the caCORRECT workflow is a quantile normalization process to align the distributions of each uploaded chip and remove global chip biases. Following this step, the

FIGURE 1. caCORRECT workflow diagram. The core workflow components include Quantile Normalization, Variance Calculation, Artifact Detection. The automated datapath runs all steps up to replacement, while the interactive datapath allows users to adjust parameters, remove problematic chip files, and update results.

variance scores are calculated to analyze data quality on a probe by probe basis. Next, image processing is run on the variance data to identify artifacts. At this point, quality metrics are calculated describing the artifact coverage and noise content of each chip and of the experiment as a whole.

After this first round is complete, additional rounds of normalization and artifact detection may follow along with calculation of updated quality scores. The purpose of this iterative scheme is to identify small to medium artifacts in later rounds that may have been previously over-shadowed by larger artifacts identified in earlier rounds. This is a common scenario as artifacts often overlap in multi-chip experiments, and artifacts influence the variance statistics by moving the calculated mean away from the true mean. Upon completion of caCORRECT, the user is presented with the following files: (1) heatmap images of all of the chips, with and without masks; (2) new versions of “clean” probe expression files with appropriate replacement applied; and (3) gene expression value tables calculated by the R implementation of RMA from data before and after caCORRECT.

caCORRECT can be run in two modes: *batch mode* or *interactive mode*. In the interactive mode, the user may browse through heatmaps of the chips in their experiment selectively removing artifacts that they encounter via a point-and-click interface. At each step of this process, the user can see how their actions are affecting the two quality metrics assigned by the

system: *artifact coverage area* and a *uniformity score*, which is derived from overall experiment variance. When the user is satisfied that the quality score cannot be improved further with additional processing, they may then elect to retrieve their microarray data files in zipped format. Figure 2 shows an example of caCORRECT’s interactive interface.

In the batch mode, the entire process of artifact removal is completely automated with no input from the user. Our experience is that batch mode gives comparable results to those produced by users during the interactive mode. Most users (including the authors), once becoming comfortable with the caCORRECT concept by using the interactive mode, elect to use the batch mode almost exclusively. In any case, if the user is unsatisfied with the results of the batch mode, they may recover their session in order to augment or touch-up these results in the interactive mode. It should be noted, however, that even advanced users find it difficult to improve results beyond those produced by the batch mode alone.

Variance Statistics

The cornerstone of any QC tool is the statistical method used for outlier detection. Here we employ a method similar to a modified *t*-statistic to describe each spot’s behavior compared to its peers on other chips. Our method is better than other methods such as absolute deviation or residuals because it reduces the

FIGURE 2. caCORRECT screen shots. These are two examples of the caCORRECT interactive mode interface under normal use. The default configuration is on the left-hand side. Clicking on the switch button (two opposing arrows) produces the image at the right-hand side with the enlarged histogram plot. The color stripes in the histogram plot correspond directly to the color bands used in the heatmap. The histogram panel can display the distributions of variance score (shown), original intensity values and normalized intensity values. As artifacts are detected, these histograms will be colored as in Fig. 3. Clicking on the magnifying glass icon will replace the histogram pane with a 5× magnification of the heatmap, with the location controlled by the mouse. This example shows the highest resolution, where 1 pixel in the large heatmap represents a spot on the array.

“ghosting” effect in which an artifact on one chip can transfer or “ghost” onto other clean chips, and confound analysis, leading to unnecessary data removal. To accomplish this, we implement a fuzzy weighting scheme that allows multiple artifacts in the same location on different chips to be observed simultaneously, while still retaining the ability to perform sequential and iterative analysis to a good stopping-point. Our method introduces a variance component, which identifies subtle artifacts in low-variance probes and ignores natural variation in high-variance probes. To begin, an artifact-weighted independent mean μ_{jk} is calculated for each location k on a chip j by using the expression values x_{ik} of that location on every other chip i in the set.

$$\mu_{jk} = \frac{\sum_{i \neq j} x_{ik} \cdot \alpha_{ik}}{\sum_{i \neq j} \alpha_{ik}}$$

The weighting factor α_{ik} equals one for spots that have not been identified as artifactual, and a small non-zero parameter λ for spots that have been identified as artifactual. Note that in the initial round of variance scoring no artifacts have yet been identified, and these procedures reduce to the commonly known sample mean, sample standard deviation, and t -statistic.

$$\alpha_{ik} = \begin{cases} 1 : & \text{not_artifact} \\ 0 \leq \lambda < 1 : & \text{artifact} \end{cases}$$

Next an artifact adjusted independent deviation score σ_{jk} is calculated in a similar manner, downplaying the weight of artifacts already identified on other chips.

$$\sigma_{jk} = \sqrt{\frac{\sum_{i \neq j} (x_{ik} - \mu_{jk})^2 \cdot \alpha_{ik}}{(\sum_{i \neq j} \alpha_{ik}) - 1}}$$

The variance statistic t_{jk} for each location on chip j is then calculated as the expression level difference from the adjusted mean, modified by the adjusted deviation.

$$t_{jk} = \frac{|x_{jk} - \mu_{jk}|}{\sigma_{jk}}$$

A non-linear scaling procedure approximate to the cumulative density function of the Student’s t -distribution is then applied to account for the statistical ramifications of the number of chips in the study, and to bound the score between 0 and 100 for consistent visualization and image processing across datasets. The result of this analysis yields a variance statistic t_{jk}^* for each spot k on each chip j in the study. High variance indicates statistical improbability, and thus artifactual tendency. Care must be taken, however, not to remove every spot in the study with high variance statistic but only to use this statistic as a guideline for identifying *regions* of chips where artifacts are present.

The variance statistic is critical to correct identification of outliers, but with this measure alone, the user does not get the critical feedback necessary to improve their microarray processing techniques. This feature is provided by incorporating the variance statistic into an intuitive heatmap based on the physical layout of the chip.

Heatmap Generation

The heatmaps implemented in caCORRECT serve a dual purpose. First, they are the critical input for both the batch-mode and user-guided image-based artifact detection process (to be described later). Second, the novel interactive nature of the heatmap in the caCORRECT interface allows the user to see and learn about the nature of the artifacts in the data, and perhaps to take steps to avoid recreating these mistakes in the future.

Heatmaps are created by thresholding the variance score described in the previous section with a default value of 80 (roughly the 80th percentile for scores). Any variance value above that threshold is automatically assigned to the “hottest” heatmap color. For a microarray chip designed using modern chip layout techniques and containing reliable data based on well-constructed (and well-executed), experimental protocols for hybridization, these spots will represent real RNA concentration differences in the sample and will be distributed randomly throughout the chip. In many cases, protocols do not achieve uniform hybridization due to uneven drying, formation of salt streaks, scratching or contamination of the microarray surface due to contact with skin or dust, miscalculated hybridization times, or failure to control environmental variables.⁷ All of these most common mistakes result in clearly visible localized patches of variance (artifacts) on the heatmap. The default threshold for batch mode is 80, but the interactive mode of caCORRECT allows the user to adjust the threshold and immediately see the resulting chip coloration. The user specifies a maximum threshold by clicking a histogram in the heatmap control panel. All chips in the experiment are represented in the histogram to help the user in identifying exceptionally high-variance chips. The available color schemes are designed to allow the user to see patterns in variance changes while minimizing bias that may result from color intensity interactions (see Fig. 3). Four resolutions are also provided for heatmaps to allow users to “blur” the data to help recognize global trends in the data.

Even the best experimental protocols can produce microarray data with varying global intensity profiles. This is especially true for microarrays produced on different days or by different labs. This difference in mean intensity must be removed before calculating the variance or producing heatmaps. We achieve this using quantile normalization as described below.

Artifact-Aware Normalization

Quantile normalization, as described in Bolstad *et al.*³ and recommended by the latest FDA MAQC

results,²² reduces noise in microarray experiment replicates by forcing the intensity distribution of each chip to be identical. A set of n distributions is from the same family of distributions (i.e., normal) if a plot of the quantiles in n dimensions results in a straight line. Projecting each quantile onto the unit diagonal vector will transform all distributions to one identical distribution. This method is generally good for the microarray problem, where the distributions are poorly defined, and parametric methods break down. However, the power of quantile normalization comes with a major caveat; if the chips are not of the same distribution, the algorithm will proceed to warp the distributions to be the same and proceed as if nothing bad happened. Fortunately, it is a reasonable assumption that high-quality microarray data from a single source on a single platform follow the same distribution. Unfortunately, this high quality is not the case in much real-world data, where chip artifacts can significantly alter the distribution of intensities on a chip. One bad chip then ends up warping the others when quantile normalization is performed, thus compromising the reproducibility of the results. We have found a way to alleviate this problem by identifying artifacts before quantile normalization, and setting them aside temporarily. In theory, perfect knowledge of artifacts would allow for perfect correction. This process is called “artifact-aware quantile normalization” and its positive effects are described in Fig. 4.

Figure 4 demonstrates the difference between classical quantile normalization and our new artifact-aware normalization using real microarray data and artifacts identified with caCORRECT’s batch mode. As can be seen in the raw intensities, the artifacts (pink) have a different distribution than the rest of the chip (blue). Once the first round of classical quantile normalization is performed, the “warping” of the non-artifacts can be seen in the way that the blue and red curves differ in shape, unlike in the artifact-aware normalization where the blue and red curves agree nicely. The noticeable hump in one of the red curves at low intensities for artifact-aware normalization is due to artifacts on another chip, and is expected in the same way that the black hump exists for the chip highlighted in this example. The original caCORRECT heatmaps for these two chips are also provided for reference (Fig. 5).

After the quantile normalization, the artifact intensities are then replaced with appropriate values as discussed in the next section. After this replacement step, the results again show that the non-artifact intensities follow a more natural distribution (the shape of the blue curve is similar to the red curve) when they are processed with the artifact-aware normalization procedure. In this case, the classical method

FIGURE 3. Heatmaps from the caCORRECT web interface. (upper left) The four color schemes: tissue, marine, sky, and lawn. (upper right) Examples of edge effect, very small splatter, and bubble effects artifacts identified in a public dataset. (lower left) The effects of changing the threshold using the slider provided on the interface to differentiate between extreme and minor artifacts. (lower right) The effects of using the four resolution options: ideal (1 pixel = 1 probe), accuracy (3x down sampling), balanced (5x down sampling), and speed (7x down sampling).

FIGURE 4. Illustration of artifact effect on the quantile normalization. Each panel is a histogram of probe intensities, where each red curve is a separate chip. One specific chip is represented with the black (all probes), blue (non-artifact probes) and magenta (artifact probes), while the other chips in the set are the red lines. Flowing from left to right are raw intensities, quantile normalized intensities, and the final intensities with outliers replaced by appropriate average values. The top two panels represent a workflow in which classical quantile normalization has been performed, while the bottom two show caCORRECT's artifact-aware normalization.

FIGURE 5. Heatmaps of artifact-rich examples. In this color scheme, dark gray regions represent artifacts, and light-blue represent clean data.

results in a sharper distribution, pushing mid-range expression values into inappropriately lower or higher value ranges.

This new artifact-aware quantile normalization scheme is implemented as the standard normalization procedure for caCORRECT. In order for it work properly however, only appropriate artifacts must be identified. If, for instance, all extremely low or high values were arbitrarily marked as artifacts, this would ruin the assumption that the remaining non-artifact intensities on a chip follow the natural distribution of chip intensities. Thus, in order to preserve the integrity of this procedure, a conservative artifact removal process is implemented as discussed in the next section.

Artifact Identification and Replacement

Perhaps the most obvious benefit of this tool is the identification and replacement of artifactual data before they can foul downstream results. caCORRECT accomplishes this by sending its own heatmaps through a battery of image processing routines which are designed to identify spatially relevant areas of high variance. The image processing routines can vary from relatively simple moving-window searches used in batch mode to complex custom-designed kernels used in interactive mode to find specific artifacts based on user input.

To maintain the specificity of artifact removal during interactive mode, we require that users provide minimal information or parameters regarding the type, size, and location of probable artifacts present on the chip image via a simple point-and-click interface. Based on the artifact type and size suggested by the user, an image kernel (mask) is generated. Simple morphological convolution is then applied to the chip heatmap image based on the custom kernel. Depending on the type of artifact being identified, various

open and close operations are then performed to further specify the artifacts. Once identified, these artifacts may be superimposed on the original chip image for visual comparison by the user. Flagged spots already indicated by proprietary software provided by Affymetrix™ encoded into the CEL file format can also be superimposed on the image for comparison during the artifact flagging process. In most cases, we have found that outliers specified in CEL files do not correspond well to our artifact definitions, but are more randomly distributed throughout the heatmap. The user has the option of retaining these outlier indexes, or replacing them upon data retrieval.

Aside from visual inspection, we suggest data quality metrics (discussed later) as a quantitative measure of the completion of the artifact removal process and resultant chip quality. The user may accept the identified artifacts at any time, or repeat the artifact removal process iteratively until they are satisfied.

Once the user is satisfied that the artifact removal is complete (or at the end of batch removal) the data may be retrieved by the user with the artifactual data appropriately replaced. The most common methods are to replace individual probe intensities with the mean or median values of that probe for each non-artifact sample in the dataset. Besides these options, caCORRECT also allows replacement with zeros, with “null,” or any other custom value that the user prefers.

Data Quality Scoring

A major goal of the caCORRECT project is not only to identify and remove errors in microarray experiments, but also to describe experimental datasets and chips according to comparable quality scores. Our efforts have produced two balanced scores: the *Uniformity Score* (described below) and *Artifact Coverage Percent*. The desired properties of a quality score for

microarray experiments are as follows: (1) If chip variation between technical or biological replicates is low, the quality score is high; (2) the quality score of an experiment composed of identical chips is maximum; (3) The quality score should be bounded for easy comparability (e.g., [0,1]); (4) Masking of high variance regions from the scoring algorithm should improve quality; and (5) the ordering of the chips should not affect the quality score.

For caCORRECT, we implemented a pairwise Normalized Cross-Correlation algorithm that satisfies all of our desired properties. This quality score serves a dual purpose. First, the score gives feedback to a lab which is generating microarray data as to the quality and repeatability of their own work. Second, it gives users of microarray data repositories some criteria by which to select a dataset. In addition to this uniformity score, we calculate an artifact coverage percentage throughout the artifact identification process for each chip, and for the dataset as a whole. This number can be used to remove highly suspect chips from the data set and improve overall data quality. It is recommended that any chip with greater than 20% artifact coverage be considered for removal from the dataset, especially if the artifact is widespread and diffuse in nature, as opposed to a sharp, localized artifact.

Generation of Test Data

In order to directly assess the quality improvement ability of caCORRECT, synthetic test data sets are generated. The test data has two advantages over real data: we have total control of which genes are differentially expressed for identifying true and false positives, and we have total control of the appearance and location of artifacts. To simulate conditions observed in real array experiments as much as possible, real values from real arrays²⁵ are used to construct the test data. Some data are harvested from obvious regions of artifact as well as from regions determined to be generally artifact free. It is worth noting that the “artifact-free” regions need not be entirely clean, because the loss of spatial information that occurs following the subsequent resampling process dilutes any artifact to negligible levels. Test microarray synthesis is performed with the following procedure:

1. Two identical templates are generated (one for each class in the biclassification problem) consisting of the numbers 1–5. The numbers are uniformly and randomly distributed across the chip, and represent the relative expression quantile for each location;
2. A specified percentage of the templates are “spiked” in random locations by altering the templates so that they are different for a location. This is to represent a differentially expressed gene;
3. Five good and bad distributions are created by first sorting, and then dividing the values extracted from a real chip into equal sized quantiles;
4. A hand drawn binary artifact mask is selected which will dictate the distribution used for each location on the generated chip (0 for bad and 1 for good); and
5. For each spot, a random number is pulled from the corresponding distribution (good or bad depending on the mask, and quantile 1–5 depending on the template).

Two datasets are created as described above, and with the following properties. Dataset 1, called the “highly separable” dataset is 640×640 , and spiked with 0.15% of the spots on the array. Of these spiked probes, the templates are always changed such that one class would be chosen from quantile 1 and the other from quantile 5. Dataset 2, called the “weakly separable” dataset is also 640×640 , but has 0.20% of its genes spiked. For the weakly separable dataset, spiking consists of altering the two classes to be in quantiles randomly selected from 2 to 4 without replacement.

To validate the effectiveness of caCORRECT in aiding biomarker discovery, two tests are performed on each simulated microarray dataset. Test one uses a student’s *t*-statistic to rank each spot in decreasing significance. The top genes identified by this ranking are then checked against the list of spiked genes, and the amount of overlap is recorded. The size of the top gene list for this test is determined by the number of spiked genes in the data so that the genes can have 100% overlap. The second method analyzing the ability of caCORRECT to aid in biomarker selection is the construction of ROC curves. For each dataset, ROC curves are created showing the rate of true positive discovery versus false positive discovery.

Testing Biomarker Reproducibility with Cross-Validation

To measure caCORRECT’s effect on the reproducibility of microarray experimental findings, ranked lists of selected biomarkers from independent datasets are compared. The goal is to find little change between ranked gene lists of two different datasets—indicating reproducibility of findings. To create these independent datasets, the dataset published by West *et al.*¹¹ is randomly split into two non-overlapping data subsets. In this way, one large 24 vs. 25 dataset is split into subsets of 12 vs. 12 and 12 vs. 13. Each of these subsets is then used as input for a support vector machines gene ranking routine twice; once without any QC, and

once after being run through the caCORRECT batch mode QC. RMAexpress was used for all gene expression calculations in this study. Figure 6 shows a schematic for this design.

The above process yields four different lists of ranked biomarkers, which are then compared in a pairwise manner to determine reproducibility of findings. First the ranks are transformed to the log 2 scale to reflect the decrease in marginal discriminatory power with an increase in rank. Second, the change in log 2 rank is then calculated between two lists for the top 100 genes on each list. This whole procedure is then repeated with different random splits to ensure repeatability.

RESULTS

Improvement in Synthetic Test Data Analysis

For both of the synthetic test datasets, caCORRECT is able to improve biomarker identification according to both the rank-overlap test, and for ROC analysis. For both datasets, the distribution of *t*-statistics for each spot becomes more significant after caCORRECT analysis, but the magnitude of the *t*-statistics increases in a more favorable manner for the spiked genes as a whole.

Analysis of spiked gene appearance in the set of top genes identified by *t*-statistic indicates that caCORRECT actually increases the amount of truly significant genes found in the top ranks of biomarker selection by *t*-statistic. Table 1 summarizes these

results, showing the increase in the percentage of genes identified that are actually spiked in during the data creation process. The remaining genes come from the same distribution, and hold no discriminating information about the two classes being investigated.

Comparing which genes are found first in the *t*-statistic ranking yields the ROC curves shown in Fig. 7. In this case, true positives are spiked in values and false-positives are the other genes which comprise over 99% of the total spots on each chip. As can be seen for both the highly separable (top) and weakly separable (bottom) ROC curves, the data that has been processed with caCORRECT (red) outperforms that of unprocessed data (blue).

Improvement in Biomarker Reproducibility

Results from the independent splitting of the West dataset are summarized in Fig. 8. It can be seen that using QC on both datasets has a significant positive effect on the reproducibility of ranked gene lists in two ways. First, distribution plots show an increase in the number of genes whose rankings do not change more than two-fold. Second, this change is enough to drag the mean of rank differences far away from the distribution that does not use QC ($p < 0.05$).

Quality Assessment of Public Data

Our artifact removal strategies are validated against real data from the scientific literature. We selected

FIGURE 6. Illustration of workflow for biomarker list comparison. Datasets consisting of probe-level microarray data are shown in light blue, expression level microarray data is in red, and ranked lists in orange. The three comparisons made here correspond to the results shown later in Fig. 8.

TABLE 1. Comparison of synthetic spiked transcript identification accuracies before and after caCORRECT.

Dataset	Total number of genes	Number of spiked transcripts	Percentage identified before caCORRECT	Percentage identified after caCORRECT
Highly separable	409,600	611	81.8%	91.5%
Weakly separable	409,600	817	55.4%	58.4%

FIGURE 7. Effect of caCORRECT on receiver operating characteristic (ROC). Curves shown are for the highly separable (top) and weakly separable (bottom) synthetic datasets. Blue is before, and red is after caCORRECT processing. Please note the difference in horizontal scales for the two panels, as the highly separable dataset needed much less false positive rates to achieve full identification of true positives. In both cases caCORRECT increased the performance of positive identification. **FIGURE 8. Distribution of differences in ranked gene lists. Normalized density plots show an increase in the number of genes that do not change rank for increasing levels of QC.**

recent experiments from ArrayExpress and Oncomine with only two constraints: each experiment must include Affymetrix CEL files and each experiment must have an associated publication.

Next we divide the chips into classes based on sample source as specified by the paper. Figure 9 shows selected heatmaps generated by caCORRECT during our survey. Our results (see Table 2) show three

interesting trends. First, the yeast genome chips produce data of overall higher quality than the human microarrays. These chips also represent the most recently acquired data. Second, in cases where artifact coverage is the greatest, the improvement in quality score due to running caCORRECT is also the greatest. Finally the quality score corresponds well to predicted similarity of the biological samples. Separating chips

FIGURE 9. Selected heatmap images generated by caCORRECT during a survey of public microarray data sets. Each heatmap is a whole microarray.

by class produces the highest quality scores when the classes are the most uniform (wild-type yeast and healthy lung tissue) while samples of low uniformity (engineered yeast and tumor samples) are generally of lower quality.

DISCUSSION

caCORRECT has been developed to aid research that uses microarrays as a primary data source. The system uses an accessible web-interface and novel

TABLE 2. caCORRECT sessions on real public data obtained by searching ArrayExpress and Oncomine.

Database	Chip model	Experiment	Initial quality	Clean quality	Chips	Artifact coverage
ArrayExpress	YG_S98	Yeast NOS stress				
		All ¹⁰	95.1495	95.8168	20	2.35%
		Wild-type only	97.3682	97.6129	12	3.60%
		Other types	92.3822	94.3465	8	4.75%
Oncomine	HU6800	Lung carcinoma				
		All ²	91.5767	93.4413	98	3.88%
		Tumor class I	90.3932	92.8775	67	4.36%
		Tumor class III	91.1865	92.8416	19	5.55%
		Healthy lung	97.7105	98.0072	10	1.64%
Oncomine	U95Av2	Breast carcinoma				
		All ¹¹	92.0755	92.5305	89	1.30%
		High-risk	92.2305	93.148	18	3.27%
		Low-risk	91.4815	92.2577	19	1.87%
		Recurring	91.483	91.8821	18	1.65%
		Non-recurring	91.799	92.4979	34	1.63%

The quality score is indicative of the actual dataset quality.

visualizations that engage users through interactivity. A novel framework for artifact detection and removal has been implemented by incorporating both user feedback and existing knowledge. caCORRECT takes a significant step beyond model-based smoothing. It targets at finding authentic artifactual instrumentation error instead of using only mathematical assumption. The results show that caCORRECT can effectively recover lost information that has been obscured by artifacts, and can increase reproducibility of biomarker selection from real microarray data. A large-scale survey of public data also demonstrates caCORRECT's ability to both assess and improve the quality of a wide array of datasets.

In conclusion, we have developed a web-based tool that is able to assess and improve quality of microarray data, which in turn will improve the quality of biomarker selection, therapeutic target studies, and pathway analysis studies in bioinformatics and systems biology. Importantly, caCORRECT improves downstream analysis of microarray data, especially in biomarker selection and translational bioinformatics. We believe that this tool could have a large impact on the wider systems biology and bioinformatics community (1) by illustrating how common the microarray artifacts are, and (2) by providing a quality assurance tool to save a tremendous amount of time and money spent in generating and utilizing experimental data.

Our future work is to evaluate artifact replacement schemes that are more sophisticated than simple median value replacement. Our preference is to not replace data in any way, but many classification schemes are difficult to adapt to data containing samples of varying dimensions.^{14,18,19,24,27} Our primary near-term objective is to implement MAGE-ML dataset import, including import of common chip platforms other than Affymetrix. Another goal is to integrate with NCI caBIG (*cancer Biomedical Informatics Grid*) by following the procedures outlined in the caCORE Software Development Kit Programmer's Guide and registering all data elements in the system with caBIG Enterprise Vocabulary Services (EVS). This would enable caCORRECT to be interoperable with the caARRAY archive, whose data model already allows for tracking of Quality Control process descriptions and metrics but lacks a standardized way to generate them.

ACKNOWLEDGMENT

This research was supported by grants from the National Institutes of Health (R01CA108468,

P20GM072069, U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to M.D.W.), and Microsoft Research eScience Fund. R.A.M. acknowledges the National Science Foundation for a graduate research fellowship. The authors are grateful to Peter Henning, Jennifer Dale, Chang F. Quo, Lauren Smalls-Mantey, Malcolm Thomas, Geoffrey Wang, Weimin Feng, Ryan Wagner, Kathy Pham as well as Drs. Georgia Chen, Scott Hemby, and John McDonald for insightful discussions. The authors would also like to thank the reviewers for their help and insight. Special recognition is due to Dr. Andrew Young for his advice and the use of his unpublished microarray data, and Deepak Sambhara and J.T. Torrance for finding and curating public datasets.

REFERENCES

- ¹Ball, C. A., I. A. B. Awad, J. Demeter, J. Gollub, J. M. Hebert, T. Hernandez-Boussard, H. Jin, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, P. O. Brown, and G. Sherlock. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* 33:D580–D582, 2005.
- ²Beer, D. G., S. L. R. Kardia, C. C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. A. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8:816–824, 2002.
- ³Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193, 2003.
- ⁴Brettschneider, J., F. Collin, B. M. Bolstad, and T. P. Speed. Quality assessment for short oligonucleotide microarray data (unpublished). Submitted to *Technometrics*, March 17, 2006.
- ⁵Brodsky, L., A. Leontovich, M. Shtutman, and E. Feinstein. Identification and handling of artifactual gene expression profiles emerging in microarray hybridization experiments. *Nucleic Acids Res.* 32:e46, 2004.
- ⁶Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32:490–495, 2002.
- ⁷Fare, T. L., E. M. Coffey, H. Y. Dai, Y. D. D. He, D. A. Kessler, K. A. Kilian, J. E. Koch, E. LeProust, M. J. Marton, M. R. Meyer, R. B. Stoughton, G. Y. Tokiwa, and Y. Q. Wang. Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* 75:4672–4675, 2003.
- ⁸Gollub, J., C. A. Ball, G. Binkley, J. Demeter, D. B. Finkelstein, J. M. Hebert, T. Hernandez-Boussard, H. Jin, M. Kaloper, J. C. Matese, M. Schroeder, P. O. Brown, D. Botstein, and G. Sherlock. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* 31:94–96, 2003.
- ⁹Holloway, A. J., R. K. van Laar, R. W. Tothill, and D. D. L. Bowtell. Options available—from start to finish—for

- obtaining data from DNA microarrays. *Nat. Genet.* 32:481–489, 2002.
- ¹⁰Horan, S., I. Bourges, and B. Meunier. Transcriptional response to nitrosative stress in *Saccharomyces cerevisiae*. *Yeast* 23:519–535, 2006.
 - ¹¹Huang, E., S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene expression predictors of breast cancer outcomes. *Lancet* 361:1590–1596, 2003.
 - ¹²Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15, 2003.
 - ¹³Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Q. Yang, S. Q. Ye, and W. Yu. Multiple-laboratory comparison of microarray platforms. *Nat. Methods* 2:345–349, 2005.
 - ¹⁴Kim, H., G. H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 21:187–198, 2005.
 - ¹⁵Kluger, Y., *et al.* Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics* 4:49, 2003.
 - ¹⁶Kuo, W. P., F. Liu, J. Trimarchi, C. Punzo, M. Lombardi, J. Sarang, M. E. Whipple, M. Maysuria, K. Serikawa, and S. Y. Lee. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat. Biotechnol.* 24:832–840, 2006.
 - ¹⁷Li, C., and W. H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A* 98:31–36, 2001.
 - ¹⁸Oba, S., M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 19:2088–2096, 2003.
 - ¹⁹Ouyang, M., W. J. Welsh, and P. Georgopoulos. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* 20:917–923, 2004.
 - ²⁰Parkinson, H., U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone, and A. Brazma. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 33:D553–D555, 2005.
 - ²¹Reimers, M., and J. N. Weinstein. Quality assessment of microarrays: visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics* 6:166, 2005.
 - ²²Shi, L., L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, and R. Peterson, *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24:1151–1161, 2006.
 - ²³Suarez-Farinas, M., A. Haider, and K. M. Wittkowski. “Harshlighting” small blemishes on microarrays. *BMC Bioinformatics* 6:294, 2005.
 - ²⁴Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525, 2001.
 - ²⁵Yu, Y. P., D. Landsittel, L. Jing, J. Nelson, B. G. Ren, L. J. Liu, C. McDonald, R. Thomas, R. Dhir, S. Finkelstein, G. Michalopoulos, M. Becich, and J. H. Luo. Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.* 22:2790–2799, 2004.
 - ²⁶Zeeberg, B. R., W. M. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narashimhan, D. W. Kang, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4:R28, 2003.
 - ²⁷Zhou, X. B., X. D. Wang, and E. R. Dougherty. Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics* 19:2302–2307, 2003.