BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 1 2011, pages 78–86
doi:10.1093/bioinformatics/btq613

Gene expression

Advance Access publication November 2, 2010

Biological impact of missing-value imputation on downstream analyses of gene expression profiles

Sunghee Oh[1,†], Dongwan D. Kang[1,†], Guy N. Brock[2,*] and George C. Tseng[1,3,4,*]

[1]Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, [2]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, [3]Department of Computational Biology and [4]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Microarray experiments frequently produce multiple missing values (MVs) due to flaws such as dust, scratches, insufficient resolution or hybridization errors on the chips. Unfortunately, many downstream algorithms require a complete data matrix. The motivation of this work is to determine the impact of MV imputation on downstream analysis, and whether ranking of imputation methods by imputation accuracy correlates well with the biological impact of the imputation.

Methods: Using eight datasets for differential expression (DE) and classification analysis and eight datasets for gene clustering, we demonstrate the biological impact of missing-value imputation on statistical downstream analyses, including three commonly employed DE methods, four classifiers and three gene-clustering methods. Correlation between the rankings of imputation methods based on three root-mean squared error (RMSE) measures and the rankings based on the downstream analysis methods was used to investigate which RMSE measure was most consistent with the biological impact measures, and which downstream analysis methods were the most sensitive to the choice of imputation procedure.

Results: DE was the most sensitive to the choice of imputation procedure, while classification was the least sensitive and clustering was intermediate between the two. The logged RMSE (LRMSE) measure had the highest correlation with the imputation rankings based on the DE results, indicating that the LRMSE is the best representative surrogate among the three RMSE-based measures. Bayesian principal component analysis and least squares adaptive appeared to be the best performing methods in the empirical downstream evaluation.

Contact: ctseng@pitt.edu; guy.brock@louisville.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 7, 2010; revised and accepted on October 28, 2010

1 INTRODUCTION

Gene expression data obtained from microarray experiments are usually peppered with missing values (MVs) that occur from a variety of reasons. Even though the technology has been improved over the past decade, MV imputation remains a necessary key step in data preprocessing. Randomly scattered MVs may be due to spotting problems, poor hybridization, inadequate resolution, fabrication errors or contaminants on the chip including scratches, dust and fingerprints. These MVs are usually flagged by absent/present calls in the Affymetrix platform, detection *P*-values in the Illumina system or other outlier detection algorithms in cDNA arrays. Since many downstream analyses require a complete dataset for implementation, MV imputation is a common practice. Many MV imputation methods have been developed in the literature (Bo *et al.*, 2004; Brock *et al.*, 2008; Kim *et al.*, 2006; Oba *et al.*, 2003; Troyanskaya *et al.*, 2001). All these methods are based on the fact that genes do not function individually, but are usually highly correlated with co-regulated genes. MV imputation methods generally belong to two categories. In the first category, expression information of a missing entry is borrowed from neighboring genes whose closeness is determined by a distance measure (e.g. correlation, Euclidean distance). Famous 'local' methods in this category include *k*-nearest neighbors (KNN; Troyanskaya *et al.*, 2001), ordinary least squares (OLS; Bo *et al.*, 2004), least squares adaptive (LSA; Bo *et al.*, 2004) and local least squares (LLS; Kim *et al.*, 2006). For the second category, dimension reduction techniques are applied to decompose the data matrix and iteratively reconstruct the missing entries. Singular value decomposition (SVD; Troyanskaya *et al.*, 2001), partial least squares (PLS; Nguyen *et al.*, 2004) and Bayesian principal component analysis (BPCA; Oba *et al.*, 2003) belong to this 'global' method category. In most methodological papers, evaluations comparing relatively few (3–5) MV imputation methods in a small number (3–5) of datasets are commonly seen. This can result in over-optimism of the newly developed algorithm, an issue that has been recently discussed in Jelizarow *et al.* (2010). The issue of whether an overall best MV imputation method exists or which MV imputation method is best suited to a given dataset was not clear until a recent comprehensive comparative study by Brock *et al.* (2008). They investigated the MV imputation performance of eight popular methods in various types of datasets (time series, multiple exposure and time series × multiple exposure) and concluded that no universally best MV imputation method exists, although three top methods (LSA, LLS and BPCA) consistently performed among the best. They further proposed an entropy-based selection scheme that predicts performance of local-based (KNN, OLS, LSA and LLS) versus global-based (PLS, SVD and BPCA) MV imputation methods in different kinds of data structures, classified by an entropy

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

measure from the distribution of eigenvalues. In general, local-based methods perform better in high complexity (larger entropy measure) datasets, while global-based methods have better performance in low complexity data (smaller entropy measure). A self-training selection scheme was also proposed to simulate from a given data and decide which MV imputation method should be used.

To date, most methodology and comparative papers mentioned above applied different variants of root mean squared error (RMSE) quantities to evaluate the performance of different MV imputation methods. The evaluation is performed as follows. A complete expression data matrix is first given, MVs are randomly generated and an imputation method is applied. Finally, the RMSEs measure the difference of the imputed values to the original true values in the complete data. The procedure is repeated for different MV imputation methods and the resulting RMSEs determine the performance of the different methods.

Although such a quantitative measure is useful to investigate the degree of data recovery by MV imputation methods, it remains unclear how the MV imputation methods affect the different downstream analyses such as biomarker detection, classification and gene cluster analysis. This is particularly important, since there is no guarantee that performance evaluations by RMSE measures are consistent with evaluations by biological impacts in downstream analyses, which is the ultimate concern in microarray data analysis. There have been a few initial efforts in this direction. Jornsten *et al.* (2005) and Scheel *et al.* (2005) investigated the impact on DE gene detection. Wang *et al.* (2006) explored in the area of classification analysis. Ouyang *et al.* (2004), de Brevern *et al.* (2004) and Tuikkala *et al.* (2008) examined gene cluster analysis. Tuikkala *et al.* (2006) studied the effect of MVs in gene ontology analysis. Each of these studies presents some partial conclusions using a smaller number of datasets, comparing limited number of MV imputation methods, and including fewer downstream analysis methods in each category, relative to the current study. Aittokallio (2009) has recently performed a comprehensive review of the MV imputation issue in large-scale studies, particularly in microarray datasets. It reviewed the biological impact evaluation papers described above, pointed out potentially partial conclusions from these smaller evaluation studies and described the importance of a future comprehensive evaluation. Our study addresses this need by providing a comprehensive and systematic evaluation to examine the biological impact of MV imputation in the three areas of downstream analyses described above: biomarker detection, classification and gene cluster analysis. Supplementary Table 1 summarizes features of the six published biological impact papers mentioned above compared with the research design in this article. To our knowledge, this is the first comprehensive evaluation study to focus on all three major downstream analyses. The result will provide deep insights into the biological impacts of MV imputation, and instruct practitioners on how to pick an MV imputation method and the expected biological impacts of imputation for a given dataset.

2 METHODS

2.1 Datasets

To perform a comprehensive comparison and evaluation, we included eight microarray datasets with a binary clinical outcome that are suitable for differentially expressed (DE) gene detection and classification analysis: GOL (Golub *et al.*, 1999), ALO (Alon *et al.*, 1999), LUO (Luo *et al.*, 2001), SIN (Singh *et al.*, 2002), BEE (Beer *et al.*, 2002), VAN (van't Veer *et al.*, 2002), LAP (Lapointe *et al.*, 2004) and YU (Yu *et al.*, 2004). We also include eight datasets with various experimental or disease conditions that are suitable for gene clustering evaluation: CAU (Causton *et al.*, 2001),

SP.AFA (Spellman *et al.*, 1998), SP.ELU (Spellman *et al.*, 1998), ALI (Alizadeh *et al.*, 2000), ROS (Hughes *et al.*, 2000), YEO (Yeoh *et al.*, 2002), BHA (Bhattacharjee *et al.*, 2001) and NCI60 (Staunton *et al.*, 2001). Supplementary Table 2 summarizes each of the datasets used in this study. Genes with zero/negative or MVs are filtered out in each dataset before being used for imputation. After filtering, the remaining genes form a complete data set (CD) that is used as input for imputation evaluation.

All Affymetrix datasets except NCI60 were imputed based on both the original unlogged data or log-transformed data (Section 2.4). NCI60 was preprocessed to generate log-transformed data as described by Culhane *et al.* (2003), and unlogged data were created by exponentiating the logged values. For the cDNA datasets, R/G values were used for the unlogged data and logged data were generated by taking logs of the ratios and were further normalized by the quantile normalization method (Bolstad *et al.* 2003). For the downstream analysis methods (DE, classification and clustering), results were primarily based on the logged datasets. However, to assess the impact of taking logarithms prior to analysis, unlogged datasets were also used for Affymetrix data.

2.2 MV imputation methods

We included eight MV imputation methods for evaluation: KNN.c, KNN.e, OLS, LSA, LLS, PLS, SVD and BPCA. The algorithms are briefly described in the Supplementary Material, 'MV imputation methods'. For simplicity, we choose parameters for each method based on our past experiences in Brock *et al.* (2008). LSA and BPCA were run using JAVA code provided by the original authors, LLS was run using the *pcaMethods* package (Stacklies *et al.*, 2007) written in R, and all other algorithms were coded in R.

2.3 Downstream analyses methods

To evaluate the biological impacts of MV imputation on downstream analyses, we consider three types of analyses commonly seen in microarray experiments: DE gene detection, classification and gene clustering. The specific methods evaluated are described below.

2.3.1 DE gene detection We included SAM (Tusher *et al.*, 2001), LIMMA (Smyth *et al.*, 2004) and *t*-test plus Benjamini–Hochberg (Benjamini *et al.*, 1995) correction (*t*-test + BH). The false discovery rate (FDR) is controlled at 5% and the default parameters are used in the packages.

2.3.2 Classification analysis We included LDA (Fisher, 1936), KNN (Fix *et al.*, 1951), PAM (Tibshirani *et al.*, 2002), and SVM (Meyer *et al.*, 2003). For LDA, KNN and SVM, we performed leave-one-out validation, selected the top $N = 5,\ 10,\ 30,\ 50,\ 100$ gene features with the largest *t*-statistics and picked the one that generates the smallest Youden index (YI) (defined as sensitivity + specificity − 1). For PAM, gene selection is embedded and we picked the threshold that generates the best accuracy. To determine the optimal K value in KNN and the kernel function in SVM, we analyzed the complete datasets (CD) for three datasets (GOL, ALO and LUO; data not shown). $K = 5$ in KNN and the linear kernel function in SVM yielded the smallest error rates, and those parameters are used throughout this article.

2.3.3 Gene clustering analysis We included K-means, SOM (Kohonen, 2001) and Mclust (Fraley and Raftery, 2002). Since the number of clusters K usually cannot be determined for a given dataset, we ran gene clustering using different choices of K. Due to the already demanding computation, we only tested $K = 5,\ 10$ and 15 for K-means, SOM and Mclust.

2.4 Quantitative evaluation: RMSE measures

Variants of RMSE are commonly used as a statistical quantity to measure the similarity of estimated values and original true values, when the original true values are known. The following simulation procedure from a complete data set (CD = $(y_{gs})_{G\times S}$) with no MVs is commonly performed in the literature, where y_{gs} is the expression intensity of gene $g(1 \le g \le G)$ and sample $s(1 \le s \le S)$. MVs are randomly generated by removing r% of values in complete data to generate data with MVs (MD). Given a MV imputation method, the missing entries in MD are imputed as $\hat{y}_{gs}$ and the imputed dataset is denoted as ID. Finally, RMSEs are used to evaluate the performance by comparing the values of missing entries in ID with those in CD.

Below, we outline six different RMSE measures that have been utilized in the literature for evaluating MV imputation methods. Bo *et al*. (2005) used a non-normalized RMSE measure between the true values and the estimated values:

$$\text{RMSE} = \sqrt{\frac{1}{\#\text{ of missing}} \sum_{\{y_{gs}\ \text{missing}\}} (\hat{y}_{gs} - y_{gs})^2}$$

Other papers normalized the RMSE measure by different normalizing constants: average value over all observations in complete data (NRMSE1; Troyanskaya *et al*., 2001), standard deviation of the values in complete data over missing entries (NRMSE2; Oba *et al*., 2003 and Kim *et al*., 2005) and root mean square of the values in complete data over missing entries (NRMSE3; Ouyang *et al*., 2004). See Supplementary Table 3 for details.

The main purpose of normalizing the RMSE is to allow for comparisons across different datasets that possibly have different intensity scales or intrinsic difficulties in MV imputation. For ranking or selecting the best MV imputation method in a given dataset, however, all the four RMSE variants provide identical results. Therefore, we will keep the NRMSE by Troyanskaya *et al*. (2001) as a representative of the four variants.

Nguyen *et al*. (2004) proposed a relative estimation error (RAE) measure to compare various imputation methods. Unlike the NRMSEs, it uses an L_1-norm and has a slight modification to alleviate drawbacks when y_{ij} equals or is close to zero:

$$\text{RAE} = \frac{1}{\#\text{ of missing}} \sum_{\{y_{gs}\text{missing}\}} \frac{|\hat{y}_{gs} - y_{gs}|}{\Phi(y_{gs})} \quad \Phi(y_{gs}) = \begin{cases} |y_{gs}| & if\ |y_{gs}| > \varepsilon \\ \varepsilon & if\ |y_{gs}| < \varepsilon \end{cases}$$

Intuitively, RAE is a better measure as it penalizes less for genes with high expression level. For example, an MV imputation error of 50 for genes with true expression level at 200 is significant, while the error of 50 becomes ignorable for genes with true expression level of 2000.

More recently, Brock *et al*. (2008) suggested the logged RMSE (LRMSE) when the expression intensities are all positive:

$$\text{LRMSE} = \sqrt{\frac{1}{\#\text{ of missing}} \sum_{\{x_{gs}\ \text{missing}\}} (\hat{x}_{gs} - x_{gs})^2},$$

where $\hat{x}_{gs} = \log(\hat{y}_{gs})$. It is easy to show that the LRMSE is an approximation of a square root of an L_2-norm version of RAE without near-zero correction (see the proof in the Supplementary Material, Proof of approximation between LRMSE and RAE-L_2):

$$\text{RAE-}L_2 = \sqrt{\frac{1}{\#\text{ of missing}} \sum_{\{y_{gs}\ \text{missing}\}} \left(\frac{\hat{y}_{gs} - y_{gs}}{y_{gs}}\right)^2}.$$

In this article, we will compare and evaluate the performance of NRMSE1, LRMSE and RAE. In our evaluations, we selected to calculate the NRMSE1 and RAE measures based on the imputations of the unlogged data to conform with the method of calculation in Troyanskaya *et al*. (2001) and Nguyen *et al*. (2004), respectively, while calculation of the LRMSE was based on imputations using logged datasets. We also compared $\exp(\hat{x}_{gs})$with the original unlogged observations, where the imputations are based on the logged data.

2.5 Biological evaluation: biological impact measures

2.5.1 Biomarker list concordance index for DE gene detection Suppose CD, MD and ID are obtained and generated according to Section 2.4. By applying a selected DE gene detection method (SAM, LIMMA or t-test + BH), one biomarker list is obtained from CD (denoted as G_{CD}) and another biomarker list can be generated from ID (denoted as G_{ID}). We define the biomarker list concordance index (BLCI) between G_{CD} and G_{ID} as

$$\text{BLCI}(G_{\text{CD}}, G_{\text{ID}}) = \frac{n(G_{\text{CD}} \cap G_{\text{ID}})}{n(G_{\text{CD}})} + \frac{n(G_{\text{CD}}^C \cap G_{\text{ID}}^C)}{n(G_{\text{CD}}^C)} - 1,$$

where $n(\bullet)$ is the number of genes in a given gene set, G_{CD}^C is the complement set of G_{CD} and G_{ID}^C is the complement of G_{ID}. Note that BLCI is equivalent of viewing the biomarker list from complete data (i.e. G_{CD}) as the gold standard and G_{ID} as the prediction result. The first term equals the sensitivity and the second term is specificity. BLCI is equivalent to the well-known YI (Youden, 1950), which is defined as the sensitivity + specificity − 1. We should note that taking the biomarker list from complete data as the gold standard is necessary since we do not know the true biomarker list of a given dataset. A high BLCI value indicates that the biomarker list from ID is similar to that from CD, and MV imputation procedure does not significantly alter the results of the downstream biomarker detection method. As a result, we expect that a good MV imputation method should generate a high BLCI value.

2.5.2 YI for classification Similarly, we utilize YI as a quantitative measure to identify the impact of MVs in classification. Since we know the true class labels of the samples in this supervised learning scenario, we can directly evaluate the YI of the prediction result from each imputed data. We expect a good MV imputation method to generate a high YI.

2.5.3 Adjusted Rand Index for gene clustering analysis The Adjusted Rand Index (ARI) (Hubert, 1985) is commonly used to evaluate the similarity between any two given clustering results. The original Rand index considers clustering relationship of any pair of objects in the data and computes the proportions of concordant pairs (two objects clustered together in both clustering or not clustered together in both clustering) among all possible pairs. The adjusted Rand index (adjusted Rand index) is a standardized version of Rand index that has expectation zero when the two clustering results are randomly generated. Similar to BLCI for DE gene detection, since we do not know the true gene clustering structure of a given dataset, we take the clustering result from CD as the gold standard and compare clustering result from ID to the gold standard by adjusted Rand index. A higher adjusted Rand index value indicates higher similarity between the two clustering results, and that the MV imputation procedure introduces a smaller impact on the downstream gene clustering analysis.

2.6 Research design and evaluation criteria

In the beginning, we considered 10 imputation methods including naïve methods of column average and row average. These two methods clearly performed poorly and were subsequently removed from consideration. Eight remaining MV imputation methods ($1 \le m \le M = 8$; KNN.e, KNN.c, SVD, OLS, PLS, LSA, LLS and BPCA) were considered, eight datasets ($1 \le d \le D_1 = 8$) for DE gene detection and classification and eight datasets ($1 \le d \le D_2 = 8$) for gene clustering were evaluated, four MV percentages ($1 \le p \le P = 4$; (r_1, r_2, r_3, r_4) = (1, 5, 10 and 20%)) were considered and finally 100 independent simulations ($1 \le n \le N = 100$) were performed. In total, $8 \times 16 \times 4 \times 100 = 51\,200$ times of random deletion from complete data matrix and then MV imputation were performed. Due to the already high-computational demand, we skipped the procedure of finding the optimal parameter for each MV imputation method in each dataset and used the optimal parameters in the comparative study by Brock *et al*. (2008). To investigate quantitative and biological criteria for deciding which MV imputation methods performed better, three RMSE measures (NRMSE, LRMSE and RAE), three DE gene detection methods (SAM, LIMMA and

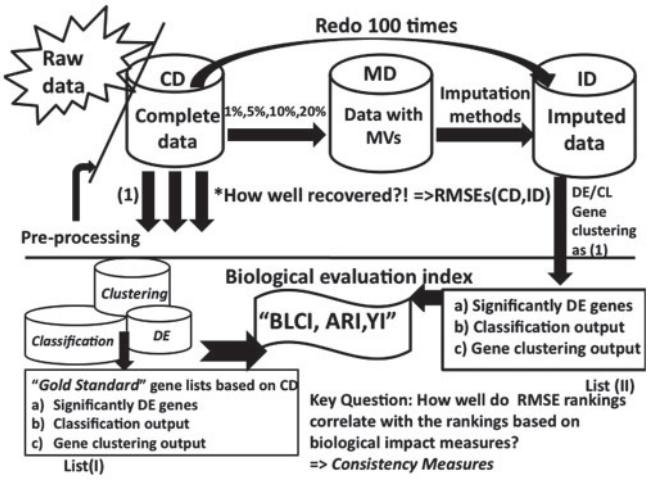

Fig. 1. Schematic illustration of the research design. Top: evaluation by RMSE measures comparing degree of recovery between complete and imputed data. Bottom: evaluation by biological evaluation indexes comparing impact in downstream analysis results. Downstream analysis results from complete data are considered 'gold standard' and downstream analysis results from imputed data are compared by BLCI, YI and adjusted Rand index.

t-test+BH), four classification methods (LDA, KNN, PAM and SVM) and three gene clustering methods (*K*-means, SOM and hierarchical clustering) were considered. In gene clustering, the number of clusters is usually not known and difficult to estimate from the data. We performed $K=5$, 10 and 15 to select the best. Therefore, we have $8\times16\times4\times100\times3=153\,600$ RMSE evaluations, $8\times8\times4\times100\times3=76\,800$ DE gene detection evaluations, $8\times8\times4\times100\times4=102\,400$ classification evaluations and $8\times8\times4\times100\times3\times3=230\,400$ gene-clustering evaluations. A schematic illustration of our research design is given in Figure 1.

2.6.1 Hypotheses and aims Our ultimate goal is to determine whether the widely used RMSE measures in the literature are adequate to decide the best MV imputation methods and how their performance compare/correlate to the actual biological impacts in different downstream analyses. To achieve this goal, we consider the following three aims:

Aim 1A: investigate whether applying different RMSE measures affects the performance ranking of the MV imputation methods.

Aim 1B: investigate whether applying different downstream analysis methods in each category (i.e. SAM, LIMMA and *t*-test+BH for DE gene detection; LDA, KNN, PAM and SVM for classification; *K*-means, SOM and hierarchical clustering for gene clustering) affects the performance ranking and selection of the MV imputation methods.

Aim 2A: if selection of the RMSE measure greatly affects the selection of the MV imputation method in Aim 1A, investigate which RMSE measure is more consistent (correlated) with the biological impact measures.

Aim 2B: quantify the consistency and correlation of the best RMSE measure (determined by Aim 2A) with the biological impact measures in terms of the performance ranking of the MV imputation methods.

Aim 3: determine which imputation methods, if any, are optimal in terms of downstream biological impact evaluation by BLCI, YI and adjusted Rand index.

2.6.2 Consistency measures To investigate the three hypotheses above, we apply Spearman's rank correlation to quantify the consistency of the performance ranking of MV imputation methods given any two evaluation measures (either RMSE measures or biological impact measures). Specifically, we define the consistency between two evaluation measures X and Y as

$$r_{dpm}^{X\times Y}=\mathrm{Cor}_{sp}((X_{1dpn},\ldots,X_{Mdpn}),(Y_{1dpn},\ldots,Y_{Mdpn})),$$

for MV imputation method m, dataset d, MV percentage r_p and simulation n. For Aim 1A and 1B, X and Y are either two RMSE measures (NRMSE, LRMSE and RAE), two BLCI measures (SAM, LIMMA and *t*-test+BH), two YI es (LDA, KNN, PAM and SVM), or two adjusted Rand indexes (*K*-means, SOM and hierarchical clustering). For Aim 2A, X is an RMSE measure and Y is a biological impact measure. We report the median consistency measure over the simulations,

$$\tilde{r}_{dp}^{X\times Y}=\text{median of }\{r_{dpm}^{X\times Y},1\le n\le N\}.$$

2.6.3 Mixed effects model For Aim 2B, we fit a mixed effects model to investigate how well the RMSE measures predict the biological impact measures. For each biological impact measure and combination of experimental conditions, we obtain a slope estimate β and a 'pseudo-R^2' value R^{*2}, which measures the proportionate reduction in the error term variance between the model which includes the RMSE measures and the model which omits them (i.e. sets β to zero). Details of the mixed effects model are given in the Supplementary Material, 'Mixed effects model'.

3 RESULTS

3.1 Aim 1—consistency among RMSE measures and among downstream analysis methods

To answer Aim 1A, Figure 2a shows the median consistency between RMSE measures, $\tilde{r}^{RMSE\times RMSE}$, in eightDE/CL datasets. In general, as the MV percentage increases (from 1% to 20%), the consistency between RMSE measures also increases. This is due to the fact that at lower MV percentages, outlying expression measurements and chance variation can have a larger influence on the RMSE measures and affect the ranking of the imputation methods, whereas at higher MV percentages this influence is abated and the measures stabilize, leading to greater agreement between the rankings. When the NRMSE and the RAE were calculated based on imputations using the unlogged datasets, the NRMSE was more correlated with the RAE than the LRMSE in almost all cases, and the consistency between the LRMSE and RAE measures is weak (Fig. 2A). But, when imputations on the logged data were used to calculate all three measures, the LRMSE generally had higher correlation with the RAE (Supplementary Tables 4[exp(log)] and 5B[exp(log)]). This is in agreement with the theoretical result that the LRMSE approximates the L_2-norm version of RAE (see Supplementary Material, Proof of approximation between LRMSE and RAE-L_2). From the variable and often low to intermediate consistency measures, we conclude that the performance ranking by the NRMSE, LRMSE and RAE for selecting the best MV imputation method greatly depends on the selection of the RMSE measure.

Similarly for Aim 1B, Figure 2B–D compares the consistency measures ($\tilde{r}^{BLCI\times BLCI}$,$\tilde{r}^{YI\times YI}$ and $\tilde{r}^{ARI\times ARI}$) of different downstream analysis methods. As with the RMSE measures, the consistency between measures increases as the MV percentage increases. For DE gene detection (Fig. 2B), the consistency between the three methods (SAM, LIMMA and *t*-test+BH) is high (around 0.62–0.97 for 20% missingness). Hence, imputation methods that perform well with respect to one DE gene detection method, in terms of the BLCI score, also perform well on the other two methods. In Figure 2C, for classification, the consistency measures are relatively low. Hence, ranking of imputation algorithms according

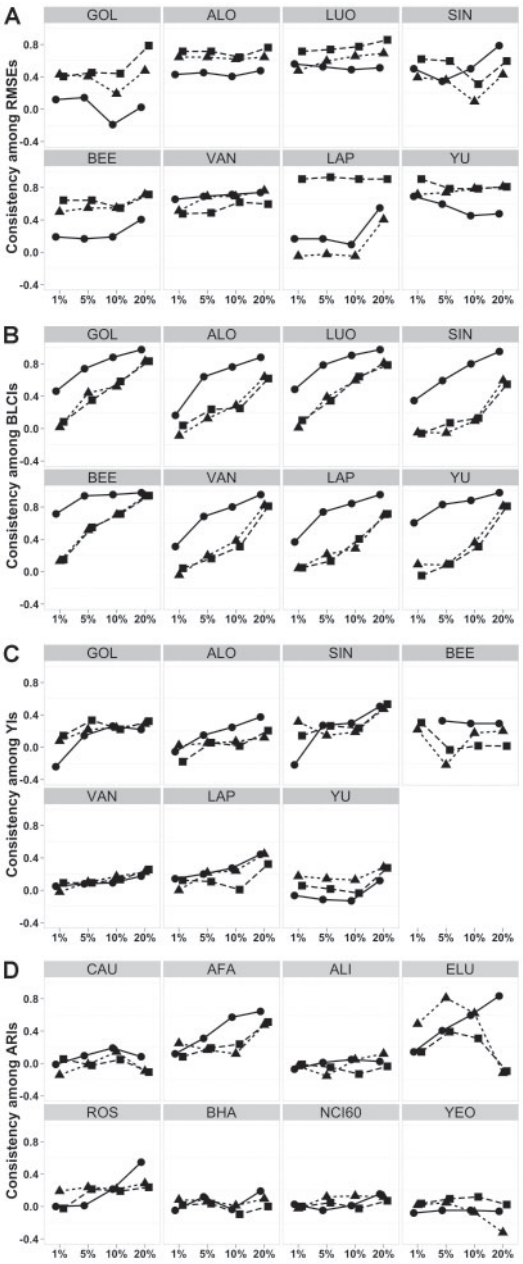

Fig. 2. Consistency measures (*y*-axis) in different MV percentage (*x*-axis). (**A**) Between three RMSE measures: circle (LRMSE versus RAE), triangle (NRMSE versus LRMSE) and square (NRMSE versus RAE). (**B**) Between three BLCIs: circle (LIMMA versus T+BH), triangle (SAM versus LIMMA) and square (SAM versus T+BH). (**C**) Between three YIs: circle (KNN versus PAM), triangle (KNN versus SVM) and square (KNN versus LDA). (**D**) Between three adjusted Rand indexes: circle (*K*-means versus SOM), triangle (SOM versus Mclust) and square (*K*-means versus Mclust).

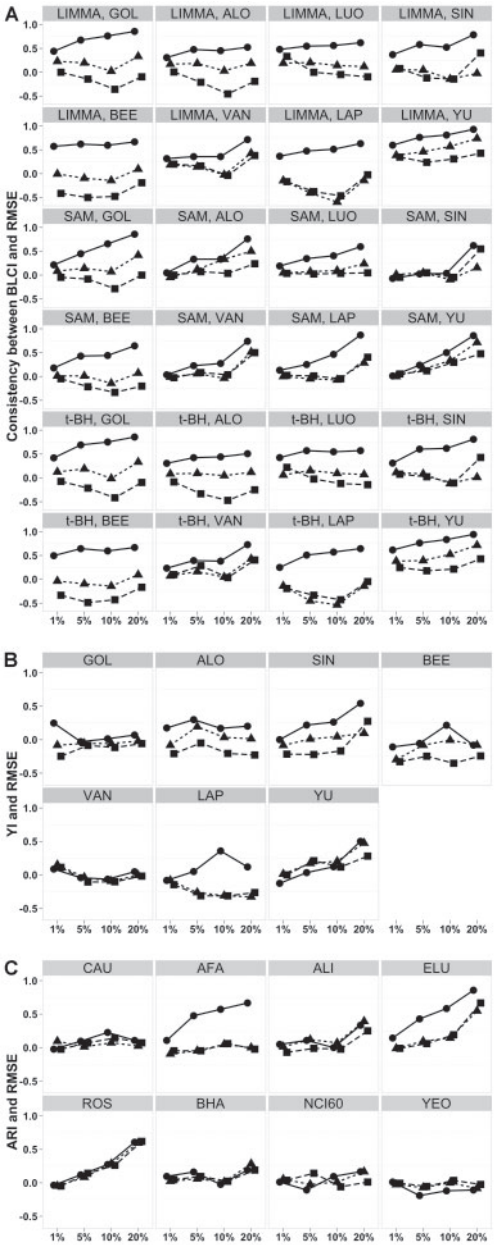

Fig. 3. Consistency measures (*y*-axis) in different MV percentages (*x*-axis). Circle (LRMSE), triangle (NRMSE) and square (RAE). (**A**) Between three RMSEs and BLCI based on LIMMA, SAM and T+BH. (**B**) Between three RMSEs and YIs based on KNN. (**C**) Between three RMSEs and adjusted Rand indexes based on *K*-means.

to one classifier is not consistent with the other three classifiers, suggesting that the impact of MV imputation on classification is more modest. In Figure 2D, for $k = 10$ clusters, the consistency among rankings of imputation algorithms based on gene clustering methods is generally as low as Figure 2C, and highly variable according to the dataset (see Supplementary Fig. 1 for $K = 5$ and $K = 15$ clusters results).

3.2 Aim 2—which RMSE measure better correlates with the biological impact measures?

Figure 3A–C shows the consistency measures between the three RMSE measures and downstream analysis methods ($\tilde{r}^{RMSE \times BLCI}$, $\tilde{r}^{RMSE \times YI}$ and $\tilde{r}^{RMSE \times ARI}$) to address Aim 2. In Figure 3A, the LRMSE measure is the most consistent with all three DE gene

detection methods (SAM, LIMMA and *t*-test+BH), followed by the NRMSE and then the RAE. The consistency increases as the MV percentage increases, ranging from 0.50 to 0.86 for the consistency between LRMSE and SAM or LIMMA at 20% missing. This indicates that performance on MV imputation tracks well with retention of the underlying DE gene list based on the complete data, so that an accurate MV imputation algorithm results in strong fidelity of the DE gene list. In Figure 3B, the consistency between the RMSE measures and YIs based on KNN classification is much lower, with the majority of values <0.4. This same pattern holds for the other classifiers as well (see Supplementary Fig. 2). Thus, the ranking of imputation algorithms on the basis of imputation accuracy does not coincide with the ranking based on classification accuracy, and the low consistency measures suggest that classification is less impacted by MV imputation compared with DE gene detection. In Figure 3C, for gene clustering, the consistency between the LRMSE and adjusted Rand index in *K*-means clustering ($K = 10$) is shown. The majority of consistency measures are around 0.2, and the LRMSE is moderately consistent with the adjusted Rand index in only two of the eight datasets (Afa and Elu). Results for 5 and 15 clusters (Supplementary Fig. 3) are roughly the same, as are the results for model-based clustering (Supplementary Fig. 4). However, for SOM, we do see a relatively high consistency between the adjusted Rand index and the LRMSE ($\geqslant 0.8$) for four out of the eight datasets (Cau, Afa, Ali and Elu), suggesting that better imputation performance also results in better preservation of clusters in this case (see Supplementary Fig. 5). Though the consistency is overall highest between the LRSME and the adjusted Rand index, due to the variable performance none of the RMSE measures can be considered a satisfactory surrogate in this case.

Figure 3A–C indicates that the LRMSE is a better quantitative measure than either the RAE or NRMSE to correlate with the biological impact measures in ranking the performance of the MV imputation methods. We further applied a complementary approach using a mixed-effects regression model to quantify the degree of consistency. The estimate of the slope term β and the corresponding pseudo-R^2 value in the mixed-effects model indicate the ability of the RMSE measure to predict the biological impact measure. Since we have concluded that the LRMSE is more consistent with the biological impact measures, we only performed the linear models for LRMSE. Intuitively, good MV imputation results in low RMSE and high biological impact measures, and we expect the β estimates to be negative and pseudo-R^2 values to be close to one. Conversely, if both β and the pseudo-R^2 are close to zero, differences of RMSE measures do not affect the biological impact measures and the selection by the RMSE is uninformative for the particular downstream analysis. In Figure 4, for regression of the BLCI on the LRMSE, we can clearly see that the slope estimates are negative in almost all situations, and the slope decreases as the MV percentage increases. At 20% missing, all slope estimates are negative and statistically significant (i.e. the 95% confidence intervals do not cover zero). Pseudo-R^2 values are also generally increasing with higher MV percentages, reaching as high as 0.8 for 20% missing. Corresponding figures for β and pseudo-R^2 values for classification and clustering are given in Supplementary Figures 6 and 7–9, respectively. In stark contrast to the figures for DE gene detection, β and pseudo-R^2 values for the regression of YI on LRMSE (classification) are all close to zero, with only a few exceptions. For gene clustering, β and pseudo-R^2 values are close to zero for both SOM and *K*-means, but association between model-based clustering (Mclust) and the LRMSE was relatively high in many cases, reaching pseudo-R^2 values of 0.8 and higher.

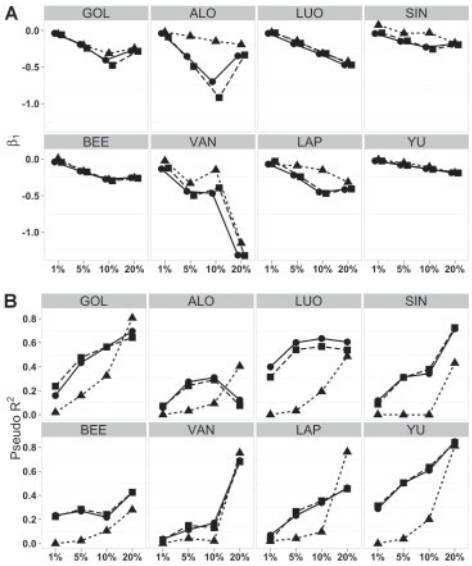

Fig. 4. β (**A**) and pseudo-R^2 (**B**) estimates from the mixed-effects model for regression of the BLCIs (circle: LIMMA, triangle: SAM, square: T+BH) on the LRMSE.

3.3 Aim 3—which imputation methods are optimal in terms of the biological impact measures (BLCI, YI and adjusted Rand index)?

To follow-up our results from Aims 1 and 2, we investigated which particular imputation methods, if any, were optimal with regards to the three biological impact measures. As shown in Figure 5A–C, averaged ranks from 100 simulations for one of the three biological impact measures, eight MV imputation methods, and two MV percentages (circle 5% and square 20%) are plotted. In Figure 5A, evaluation by BLCI clearly determines LSA and BPCA as the top performing MV imputation methods in all eight datasets. For YI evaluation in Figure 5B, the trend is not clear. Figure 5C of adjusted Rand index evaluation generates results in between that of BLCI and YI, in that LSA and BPCA show slight evidence of better performance. Figure 6 summarizes Figure 5A–C by demonstrating the distribution of averaged ranks in eight datasets and different MV percentages using boxplots. Again, DE gene analysis by BLCI evaluation shows strong evidence of best performance by LSA and BPCA methods (Fig. 6A), while classification analysis by YI evaluation shows no preference (Fig. 6B). When comparing 5 and 20% MV percentages, 20% missingness shows stronger discrepancy of averaged rankings across the eight MV imputation methods, a trend repeatedly observed in Aims 1 and 2. For gene clustering, the effects of imputation are intermediate between that of DE gene detection and classification. In gene-clustering methods, ranking of MV methods in terms of the adjusted Rand index largely depends on the choice of clustering method and the number of clusters, as shown in Supplementary Figures 11–13. And, in some cases, poor-performing methods based on the LRMSE perform well based on the adjusted Rand index (e.g. SVD for the BHA and YEO

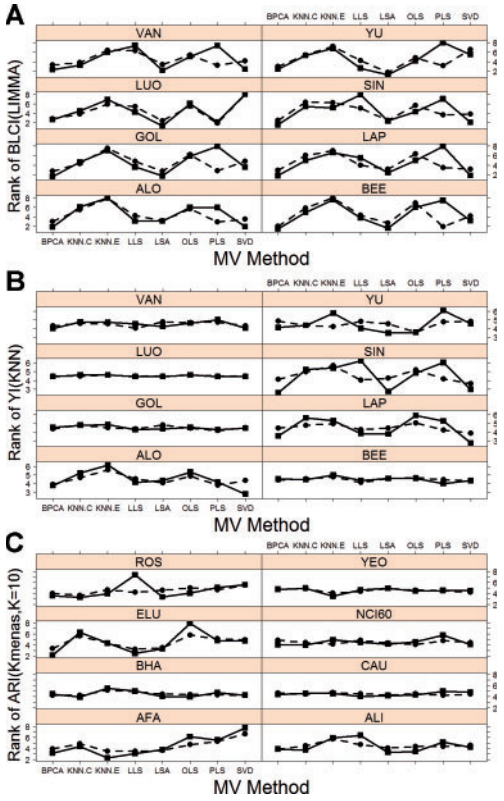

Fig. 5. Selection of the best MV imputation methods from biological impact analysis. (**A**) Averaged ranking of the eight MV imputation methods from BLCI measure by LIMMA in eight datasets. (**B**) Averaged ranking from YI by KNN. (**C**) Averaged ranking from adjusted Rand indexes by K-means, $K = 10$. Circle: 5% MV percentage. Square: 20% MV percentage.

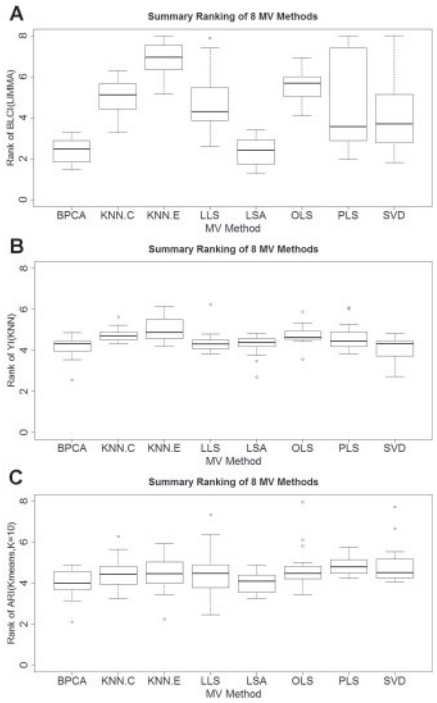

Fig. 6. Summary from Figure 5A-C. Distribution of averaged ranks of the eight datasets and different MV percentages are summarized by boxplots. Lower ranks represent better MV imputation methods judged by the biological impact measures. (**A**) Boxplot for ranking of 8 MV methods from BLCI measure by LIMMA. (**B**) Boxplot for ranking of 8 MV methods from YI by KNN Classifier. (**C**) Boxplot for ranking 8 MV methods from adjusted Rand Indexes (ARI) by K-means ($K = 10$).

datasets). Overall, it is difficult to predict which imputation method is optimal for clustering analysis on the basis of imputation accuracy, so that assessment of the downstream results becomes necessary to guarantee an optimal choice.

4 DISCUSSION AND CONCLUSIONS

Our large-scale comparative study confirms that MV imputation methods for genomic data should be examined in terms of their impact on commonly performed downstream analyses. The main novelty in this study was to investigate the correlation between RMSE-based measures, which has been the most commonly employed measures for selection and ordering of MV methods so far and measures of the biological impact on downstream analysis including DE gene detection, classification and clustering. Although several prior studies have investigated the impact of MV imputation on these downstream analyses individually, to our knowledge our study is the first to systematically evaluate the impact of MV imputation on all three of these areas of downstream analysis, using a large variety of datasets. In the following paragraphs, we highlight our main conclusions regarding the impact of MV imputation on each of the three main downstream analysis areas, and contrast our results with other recent literature.

Our investigation of different RMSE-based measures revealed that the consistency between rankings of MV imputation measures based on these measures is not high. This suggests that the selection of RMSE-based measures for evaluation of MV methods should be taken more carefully, an issue that has not been addressed in previous studies. Based on the results, the overall agreement between the three different RMSE measures (RAE, NRMSE and LRMSE) was generally moderate to high, though it varied from dataset to dataset and in some cases the consistency between measures was surprisingly weak (e.g. LRMSE versus NRMSE on the LAP dataset, see Fig. 2A). This suggests that choice of RMSE-based measure can give different conclusions regarding the selection of MV methods. The agreement between rankings increased as the MV percentage increased, which is reasonable since the separation between imputation methods should be greater with higher percentages of MVs.

In evaluating the biological impact of MV imputation on downstream analyses commonly carried out after estimation of MVs, we found that detection of DE genes was the most sensitive analysis to the choice of imputation method, while classification was the least sensitive and gene clustering was intermediately affected. Previous studies that investigated the impact of MV imputation on DE gene detection include Jornsten *et al*. (2005) and Scheel *et al*. (2005). Jornsten *et al*. investigated six methods [SVD, BPCA, KNN, Gaussian mixture clustering (GMC), row imputation and LinCmb] and three datasets, while Scheel *et al*. only investigated two methods (LinImp and KNN) and two datasets. In both cases, the more sophisticated imputation methods (BPCA, GMC, LinCmb

and LinImp) outperformed the simpler KNN and SVD methods. This is in agreement with our results, as we found that in general two of the three top-performing methods from Brock *et al*. (2008) (BPCA and LSA) also performed well for DE gene detection. While some of the methods evaluated in Jornsten *et al*. (2005) and Scheel *et al*. (2005) may outperform the methods we investigated, the important discovery from our study is that choice of imputation algorithm has an important impact on DE gene discovery, and that further this impact tracks well with the performance of the imputation algorithms in terms of the LRMSE. The strength of the correlation between the LRMSE and the biological impact ranking for DE gene detection (BLCI) was appreciably high, with pseudo-R^2 values as high as 0.8.

In contrast, for classification analysis the consistency between rankings of imputation methods based on RMSE measures and biological impact measures (YI) was low (around 0 in most situations), suggesting less biological impact of MV imputation on classification. This conclusion is consistent with Wang *et al*. (2006), who found that KNN, LLS and BPCA all had relatively similar performance in terms of classification accuracy for the five datasets they evaluated. This is intuitively reasonable because classification methods are analyses with contributions from multiple genes, and are thus more 'robust' to variation in MV estimation.

While DE gene detection and classification presented relatively clear-cut cases where MV imputation had significant and little impact, respectively, the impact of imputation on clustering analysis was more mixed. Consistency between RMSE measures and the adjusted Rand index for clustering was a mixed bag, and tended to depend both on the dataset and the number of clusters. Other recent studies that investigated the impact of MV imputation on clustering analysis include Tuikkala *et al*. (2008) and Celton *et al*. (2010). Tuikalla *et al*. evaluated eight datasets using *K*-means clustering, and found that significant differences between imputation methods in terms of RMSE measures did not translate into significant differences in terms of the original gene clusters or biological interpretations. Celton *et al.* (2010) evaluated the effect of MV imputation on both *K*-means and hierarchical clustering, using five different datasets and 12 imputation methods. Like Tuikkala *et al*., they found that *K*-means was relatively robust to MVs in terms of conserving gene associations, while hierarchical clustering was more sensitive. Our results agree with these assessments, in that *K*-means clustering seems little impacted by MV imputation, while other methods (SOM) appear to be more affected. One surprising result is that they found the EM_array method by Bo *et al.* (2004) to be the optimal method based on both the RMSE and the clustering assignment index, which was not evaluated in our study as it is a less technically advanced method compared with LSA. While BPCA, LSA and LLS frequently did well in regards to the clustering biological impact, there was definitely no single best imputation method in this regard.

Our selection of biological impact measures for the effect of MV imputation on downstream microarray analyses is motivated by choosing a measure that is both comprehensive and intuitive. The BLCI and YI were selected because they capture both the sensitivity and specificity of the result in a single measure, and the adjusted Rand index is a well-known and widely used measure of concordance between two clustering partitions. However, other measures of the impact of imputation on downstream analysis are certainly justifiable, and we feel the main conclusions from these studies are still comparable with our study. One potential limitation of the BLCI is that for smaller missing percentages, the BLCI score may be strongly impacted by genes without MVs compared with those with MVs, and we did not evaluate these separately. However, for higher MV percentages the majority of genes will have at least one MV, so by evaluating a wide range of missing percentages we feel that addresses this issue indirectly. Another limitation is that the false discovery rate was not evaluated separately, and recent studies have shown that subsamples of expression matrices can produce gene lists with low FDRs (Zhang *et al*., 2008). One last point is our use of the term 'biological' impact. In our sense, 'biological impact' relates to the impact on downstream analysis methods, which we feel is of direct concern to biologists because these methods are used to select the genes and biomarkers that form the basis for the biological interpretation of the study. However, our use of the term is different from studies that measure MV imputation performance on the basis of external information from biological databases, e.g. GO terminology (Sehgal *et al*., 2008; Tuikkala *et al*., 2008).

While the primary analysis in our study is based on the downstream analysis of the logged data, typical of practice, we also examined the biological effects on unlogged data to avoid favoritism toward the LRMSE measure. While the consistency between the BLCI- and RMSE-based rankings is still relatively high, the LRMSE has the highest correlation only for one-half of the datasets, while the NRMSE has the highest correlation for the other half (Supplementary Table 6). Consistency between RMSE and YI is still very low for classification using the unlogged data (data not shown). For gene clustering, both logged and unlogged data were inconsistent in terms of which RMSE measure was the most consistent with imputation rankings based on the adjusted Rand index, and no RMSE measure was truly adequate (data not shown). Again, though there is no considerable difference in consistency results based on logged versus unlogged data, we suggest taking log transformations prior to imputation as well as downstream data analyses, due to the potential undue influence of outliers for unlogged data (Eisen *et al*., 1999; Kerr *et al*., 2000).

We conclude by highlighting the main take-home messages from our study. Prior to deciding which imputation algorithm to use for MVs in microarray data, it is informative for investigators to know which areas of downstream analysis are even impacted by MV imputation. To this end, we have shown that while choice of imputation method has a strong impact on the results of DE gene detection, it has relatively little impact on classification accuracy and an intermediate affect on clustering results. Further, for differential expression methods, the LRMSE may serve as a representative surrogate for selecting the optimal imputation method, so that MV selection methods like those presented in Brock *et al*. (2008) should provide a good choice of imputation algorithm for DE gene analyses. However, for clustering analysis, no single RMSE truly suffices, so that evaluation and comparison of MVs should be explored directly in terms of biological impact measures. While the more sophisticated imputation methods (BPCA and LSA) were generally the top performers in terms of LRMSE and DE gene detection (BLCI), for classification accuracy there was no uniformly best method and for gene clustering some algorithms that performed poorly in terms of LRMSE (SVD) did well in terms of the clustering biological impact measure (adjusted Rand index). Hence, there is room for further sophisticated investigation of imputation

performance based on data complexity and other characteristics of the dataset (Brock *et al*., 2008), which will be the basis for future studies.

ACKNOWLEDGEMENTS

The authors graciously acknowledge the support of the University of Louisville Cardinal Research Cluster and the expert assistance of Harrison Simrall in submitting the simulation studies.

Funding: National Institutes of Health (KL2 RR024154-03 to G.C.T., partially); University of Pittsburgh (Central Research Development Fund, CRDF; Competitive Medical Research Fund, CMRF, to G.C.T. partially); Department of Energy (grant 10EM00542 to G.N.B. partially) and National Institutes of Health (grants P30ES014443, P20RR017702 and RC2AA019385-01).

Conflict of Interest: none declared.

REFERENCES

Aittokallio,T. (2010) Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief. Bioinformatics*, **2**, 253–264.

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Alon,U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Bhattacharjee,A. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 790–795.

Beer,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Bo,T.H. *et al.* (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res*, **32**, 1–8.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and varicance. *Bioinformatics*, **19**, 185–193.

Brock,G.N. *et al.* (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*, **9**, 1–12.

Causton,H.C. *et al.* (2001) Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell*, **12**, 323–337.

Celton,M. *et al.* (2010) Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC *Genomics*, **11**, 1–16.

Culhane,A.C. *et al.* (2003) Cross-platform comparison and visualization of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 1–15.

de Brevern,A.G. *et al.* (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**, 1–12.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Fisher,R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann Eugen.,* **7**, 179–188.

Fix,E. and Hodges,J.L (1951) Discriminatory analysis, nonparametric discrimination: consistency properties. *Technical Report 4*, USAF School of Aviation Medicine, Randolph Field, Texas.

Fraley,C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hubert,L. and Arabie,P.(1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Hughes,T.R. *et al.* (2000) Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet*., **25**, 333–337.

Kerr,M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Jelizarow,M. *et al.* (2010) Over-optimism in bioinformatics: an illustration. *Bioinformatics*, **21**, 1990–1998.

Jornsten,R. *et al.* (2005) DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, **21**, 4155–4161.

Kim,H. *et al.* (2006) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **22**, 1410–1411.

Kohonen,T. (2001) Self-organizing maps of massive databases. *Eng. Intell. Syst. Elect. Eng. Commun.*, **9**, 179–185.

Lapointe,J. *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 811–816.

Luo,J. *et al.* (2001) Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. *Cancer Res.*, **61**, 4683–4688.

Meyer,D. *et al.* (2003) The support vector machine under test. *Neruocomputing*, **55**, 169–186.

Nguyen,D.V. *et al.* (2004) Evaluation of missing value estimation for microarray data. *J. Data Sci.*, **2**, 347–370.

Oba,S. *et al.* (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Ouyang,M. *et al.* (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**, 917–923.

Scheel,I. *et al.* (2005) The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, **21**, 4272–4279.

Sehgal,M.S. *et al*. (2008) Ameliorative missing value imputation for robust biological knowledge inference. *J. Biomed. Inform.*, **41**,499–514.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.,* **3**, Article3.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Stacklies,W. *et al.* (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Staunton,J.E. *et al*. (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA*, **98**, 10787–10792.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tuikkala,J. *et al.* (2006) Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, **22**, 566–572.

Tuikkala,J. *et al.* (2008) Missing value imputation improves clustering and interpretation of gene epxression microarray data. *BMC Bioinformatics*, **9**, 1–14.

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

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

Wang,D. *et al.* (2006) Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. *Bioinformatics*, **22**, 2883–2889.

Yeoh,E. *et al*. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.

Youden,W.J. (1950) Index for rating diagnostic tests, *Cancer*, **3**, 32–35.

Yu,Y.P. *et al.* (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790–2799.

Zhang,M. *et al.* (2008) Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics* **24**,2057–2063.