Title

Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices

Journal

Briefings in Bioinformatics, 18(2)

ISSN

1467-5463

Authors

Taylor, Sandra L
Ruhaak, L Renee
Kelly, Karen
et al.

Publication Date

2017-03-01

DOI

10.1093/bib/bbw010

Peer reviewed

OXFORD

Briefings in Bioinformatics, 18(2), 2017, 312–320

doi: 10.1093/bib/bbw010
Advance Access Publication Date: 19 February 2016
Paper

Effects of imputation on correlation: implications for analysis of mass spectrometry data from multiple biological matrices

Sandra L. Taylor, L. Renee Ruhaak, Karen Kelly, Robert H. Weiss and Kyoungmi Kim

Corresponding author. Sandra L. Taylor, Department of Public Health Sciences, University of California, Davis, CA 95616, USA. Tel.: (916) 703-9171; Fax: (916) 703-9124; E-mail: sltaylor@ucdavis.edu

Abstract

With expanded access to, and decreased costs of, mass spectrometry, investigators are collecting and analyzing multiple biological matrices from the same subject such as serum, plasma, tissue and urine to enhance biomarker discoveries, understanding of disease processes and identification of therapeutic targets. Commonly, each biological matrix is analyzed separately, but multivariate methods such as MANOVAs that combine information from multiple biological matrices are potentially more powerful. However, mass spectrometric data typically contain large amounts of missing values, and imputation is often used to create complete data sets for analysis. The effects of imputation on multiple biological matrix analyses have not been studied. We investigated the effects of seven imputation methods (half minimum substitution, mean substitution, *k*-nearest neighbors, local least squares regression, Bayesian principal components analysis, singular value decomposition and random forest), on the within-subject correlation of compounds between biological matrices and its consequences on MANOVA results. Through analysis of three real omics data sets and simulation studies, we found the amount of missing data and imputation method to substantially change the between-matrix correlation structure. The magnitude of the correlations was generally reduced in imputed data sets, and this effect increased with the amount of missing data. Significant results from MANOVA testing also were substantially affected. In particular, the number of false positives increased with the level of missing data for all imputation methods. No one imputation method was universally the best, but the simple substitution methods (Half Minimum and Mean) consistently performed poorly.

Key words: mass spectrometry; missing data; imputation; multivariate analysis; within-subject correlation; metabolomics

Introduction

Advances in and declining costs of mass spectrometry have supported an increase in the number of mass spectrometric (MS) omics studies. As part of this increase, some investigators are collecting and analyzing multiple biospecimen types (herein referred to as 'matrices') from participating subjects such as serum, plasma, tissue and urine for multi-domain assessment [1–3]. By investigating multiple biological matrices simultaneously, investigators hope to enhance discovery of promising

Sandra L. Taylor is a Principal Statistician with the Department of Public Health Sciences, Division of Biostatistics, at the University of California, Davis, CA, USA.
L. Renee Ruhaak was a post-Doctoral Researcher at the Department of Chemistry at UC Davis. She is currently an Assistant Professor at MD Anderson Cancer Center, Houston, TX.
Karen Kelly is a Professor of Medicine at the University of California, Davis, Associate Director for Clinical Research at UC Davis Comprehensive Cancer Center and the Jennifer Rene Harmon Tegley and Elizabeth Erica Harmon Endowed Chair for Cancer Clinical Research.
Robert H. Weiss is Chief of Nephrology at the Sacramento VA Medical Center and a Professor of Medicine with the University of California, Davis, CA, USA.
Kyoungmi Kim is an Associate Professor with the Department of Public Health Sciences, Division of Biostatistics, at the University of California, Davis, CA, USA.
Submitted: 8 October 2015; **Received (in revised form):** 14 January 2016

biomarkers for disease, understanding of disease processes and mechanisms and identification of therapeutic targets.

In analyzing multi-matrix omics data from the same subjects, investigators typically analyze each matrix individually and then qualitatively compare results across the matrices [3–6]. However, this approach does not take advantage of the inherent correlation between biological matrices collected from the same subject. Use of multivariate methods such as multivariate analysis of variance (MANOVA) has the potential to increase power to reveal significant results over independent analyses. Further, multivariate methods could preserve power by avoiding adjusting for multiple testing across matrices, although many authors restrict such adjustments to each matrix.

One challenge in analyzing MS data using multivariate methods is the large amount of missing data [7, 8] because most statistical procedures require complete data. Various strategies are available for handling missing data [8, 9]. A common approach is to use a combination of dropping compounds with missing values above a predetermined percentage and imputing any remaining missing values for the compounds retained for statistical analysis. A wide variety of imputation methods have been developed and extensively evaluated particularly in the microarray literature [10].

MS data are similar to microarray data in that both yield quantitative measures of many individual compounds from a single sample and are commonly analyzed with similar techniques. However, there are marked differences in the data-generating processes of these two data types and important differences with respect to missing values. First, MS data typically has a much larger amount of missing data than microarray data. Missing values commonly account for <10% of microarray data, but in MS studies, 20–50% missing values are common [11, 12]. Imputation procedures that perform well with small amounts of missing data might not perform favorably with larger amounts. Second, and more importantly, the missing data mechanisms differ between the two technologies. In microarray studies, missing values arise owing to a variety of technical problems with no one mechanism dominating [10, 13]. The patterns of missing values in microarray data have been largely considered to be missing completely at random or missing at random [13] although Scheel noted that some missingness in microarray data can be nonrandom, occurring when a signal is too low or irregular to distinguish it from background [14]. In MS data, missing patterns have been shown to be strongly nonrandom owing to markedly increased missingness with declining compound abundance [7, 11]. As a result of these differences, conclusions on the performance of imputation methods based on microarray data investigations might not adequately capture performance when applied to MS data, and understanding the performance of imputation methods in the context of MS data is critical.

Recently, several studies have investigated imputation for MS data [7, 15, 16]. These studies have focused on the degree of deviation between true and imputed values [7, 11] and the subsequent impacts on significance testing [7, 16] or the impact of imputation approaches on within-matrix multivariate analyses such as partial least squares regression or principal components analysis [7, 15, 17]. However, imputation could also change the correlation structure of the data, and for studies involving multiple biological matrices, affect the between-matrix correlation for a compound. Acceptable performance of imputation methods within a biological matrix in terms of normalized root mean squared error, identification of differentially regulated compounds, or classification cannot be assumed to preserve the between-matrix correlation. No previous studies have investigated the impact of imputation methods on the correlation of compounds quantified using MS between biological matrices. Because the results of multivariate statistical methods that integrate results from multiple matrices such as MANOVA are influenced by the between-matrix correlation structure [18], and with multi-matrix omics studies becoming more common, it is important to understand the impact of imputation on the between-matrix correlation and its consequence in inferential testing. Therefore, we investigated the effects of imputation of (mainly not random) missing values on intra-subject correlation between different matrices in multivariate analysis using simulations with different data configurations and real MS omics data.

Methods

We used three real MS omics data sets to assess the effect of three imputation methods on the between-matrix correlation structure and MANOVA results. Characteristics of the three real data sets are described below and summarized in Table 1. We then conducted a simulation study to further understand the effect of imputation under a range of missingness for data sets with known correlation structures and effect sizes.

Real data sets

Renal cell carcinoma xenograft metabolomics

A metabolomics study of tissue, serum and urine was conducted of renal cell carcinoma using xenograft and sham surgery control mice. Human Caki-1 cells were xenografted into seven nude mice and seven mice were subjected to sham surgery. All mice were sacrificed 34 days after surgery when the xenografted animals became moribund. Terminal serum was collected, and tumor (from xenografted animals) and normal kidneys (from sham surgery animals) were removed for tissue analysis. Urine was collected 32 days after surgery (2 days before sacrifice). Nontargeted metabolomics was accomplished using three different platforms: ultra-high performance liquid chromatography/tandem mass spectrometry (UHLC/MS/MS2) optimized for basic species, UHLC/MS/MS2 optimized for acidic species and gas chromatography/mass spectrometry. Urinary metabolite values were creatinine normalized to account for urine concentration differences among samples, and tissue samples were normalized to equal mass before chromatographic analysis. Detailed information on the experimental procedures and metabolomic platforms, including sample extraction process, instrumentation configurations and conditions and software approaches for data handling, were previously described in detail [4].

Lung cancer serum and plasma glycomics

Blood samples (serum and plasma) were collected from 43 subjects diagnosed with non-small-cell lung cancer (NSCLC) adenocarcinoma and 43 healthy controls recruited during a 4-year period (2010–14) from the UC Davis Medical Center and Cancer Center Clinics. Cancer patients were frequency matched with controls for gender, age and smoking history. Glycomics analysis was performed by enzymatic release of the N-glycans followed by purification and subsequent analysis using an Agilent nanoLC coupled to an Agilent time-of-flight mass spectrometer (nLC-TOF-MS) equipped with a Chip-cube. Glycan separation was performed using a porous graphitized carbon stationary phase on a chip. Further information on sample processing, the

Table 1: Characteristics of real mass spectrometry omics data sets evaluated

Data set	Biological matrices	N	Number of compounds detected in at least one matrix	Number of compounds detected in all matrices	Number of compounds detected in all samples and matrices
Kidney Cancer Xenograft	Tissue, Serum, Urine	14	485	100	60
Lung Cancer Serum/Plasma Glycomics	Serum, Plasma	86	331	330	51
Lung Cancer Blood Glycomics	Serum, Plasma, Dried Blood	20	323	323	27

MS analysis and additional data processing is contained in [19, 20].

Lung cancer blood glycomics

Serum, plasma and dried blood spots were obtained from 10 subjects with NSCLC adenocarcinoma and 10 healthy controls matched for gender, age and smoking history. Glycomics analysis was performed using nLC-TOF-MS as described in the previous section. Collection, processing and analysis of serum and plasma samples is described in more detail in [19, 20] and in [21] for the dried blood samples.

Inducing and imputing missing values

To prevent introduction of undesired artifacts in this study, we restricted these data sets to compounds detected in all samples in all biological matrices (Table 1) and then created data sets with 10%, 25% and 50% missing values, a range reflecting our real data sets (Supplementary Table S1). We generated missing values through a restricted random selection procedure similar to [14]. In our approach, values below a predetermined threshold were randomly selected and set to missing. Specifically, to generate 10% total missing, intensity values below the 25th quantile of each biological matrix were randomly selected and set to missing. To generate 25% missing, we sampled below the 50th quantile and for 50% missing we sampled from values below the 75th quantile. This approach yielded patterns of missingness similar to the nature of real data sets where the amount of missing values tends to increase with declining intensity [7, 16] and some compounds have no missing values as previously shown in [19] from the experiments by our group (Supplementary Figure S1).

We imputed missing values using seven methods that have been used in omics studies: (1) substituting one-half the minimum compound-specific value (Half Minimum), (2) mean substitution where we substituted the missing value with the cancer group-specific mean of observed values for the compound (Mean), (3) k-nearest neighbor (KNN) [10], (4) local least squares regression (LLS) [22], (5) Bayesian principal component analysis (BPCA) [23], (6) singular value decomposition (SVD) [10] and (7) random forest (RF) [24]. These methods encompassed a range of approaches including simple substitution methods (Half Minimum, Mean), local similarity methods (KNN, LLS, RF) and global structure methods (BPCA, SVD). KNN imputation was conducted using the impute package in R [25] using 10 neighbors and with the maximum level of missing allowed for a compound before mean imputation set at 80%. For LLS, we used the llsImpute function from the pcaMethods package in R [26]. Spearman's correlation was used to identify the five nearest neighbors using all compounds. Compounds were centered to a mean of 0. BPCA and SVD imputation methods were implemented using the pca function in the pcaMethods package in R. For both methods, three components (principal components or latent variables) were used and compounds were centered and scaled to unit variance. We used the missForest function in the missForest R package to conduct RF imputation [27].

Analysis of real data sets

For the three real data sets, we induced 10%, 25% and 50% missing values and imputed the missing values using the seven methods. We first compared the between-matrix correlations of the full and imputed data sets. Then, for each compound, we tested for differences between cancer and control subjects using a MANOVA and compared results from the full and imputed data sets. We defined true positives as the number of the significant results ($P < 0.05$) from the full data set that were also significant with the imputed data set and false positives as the number of significant results in the imputed data set that were not significant in the full data set. The numbers of true and false negatives also were calculated. Intensity values were $\log_2$ transformed for all analyses to accommodate the underlying assumption of MANOVA. In addition to the MANOVAs, we used *t*-tests to conduct single matrix tests using non-missing values. We identified a compound as significantly differentially regulated if the *t*-test was significant in any matrix based on a Bonferroni adjusted threshold.

Simulation study

A simulation study was conducted to measure the effects of imputation on the between-matrix correlation and MANOVA results under known conditions. For our correlation investigations, we simulated 30 samples from a bivariate standard normal distribution with correlation of ± 0.75, ± 0.5, ± 0.25 and 0. These values were exponentiated to yield log normal distributions. One thousand data sets, each consisting of 100 compounds, were simulated. Each data set represented a single experiment that identified 100 compounds in two biological matrices from the same subject. Missingness levels of 10%, 25% and 50% were induced in the same manner as for the real data sets and imputed with the seven methods. We graphically compared correlations of the full data set to the imputed data sets.

For our MANOVA investigations, we simulated correlated data for two biological matrices from the same subject with subjects in two conditions (e.g. cases versus controls). We simulated data sets consisting of 15 samples per group from a bivariate normal distribution with variances of 1 and between-matrix correlations of ± 0.75, ± 0.5, ± 0.25 and 0. 'Controls' were always set to have a mean 0; 'Cases' had (1) a mean of 0.5 in both biological matrices or (2) in only one of the biological matrices. Missingness levels of 10%, 25% and 50% were induced as before and the bivariate normal values were exponentiated to yield log normal distributions. We simulated 1000 data sets,

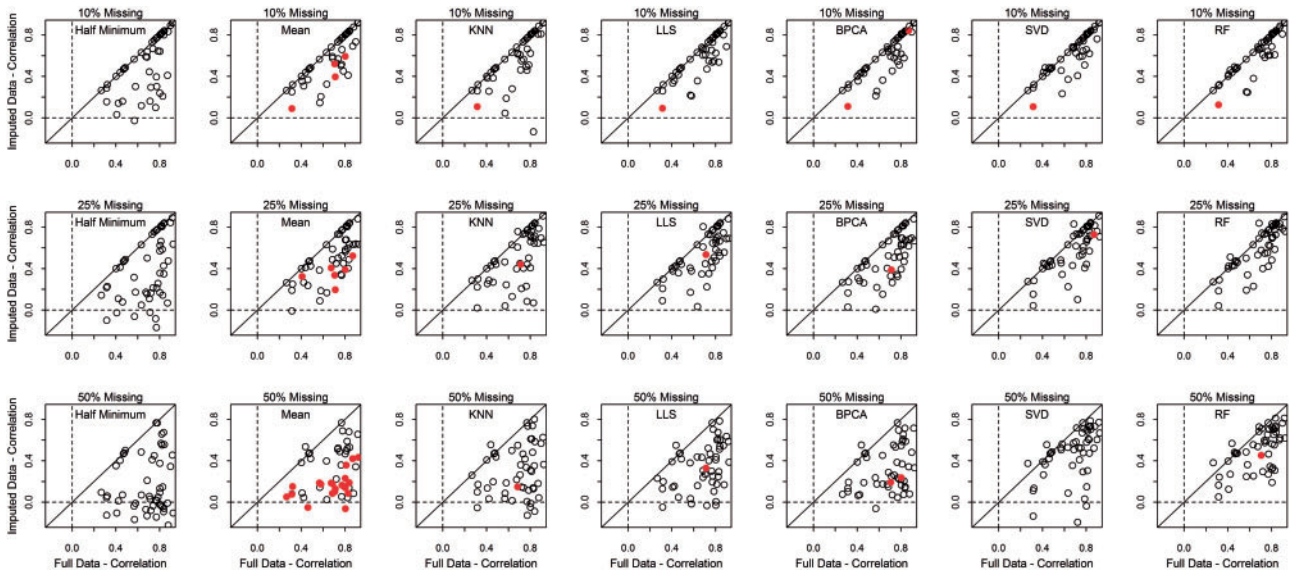

Figure 1: Comparison of between-matrix correlation of intensity values of glycans measured in plasma and serum of lung cancer patients for the full data set and imputed data sets with 10%, 25% and 50% missing values after imputation using seven methods (Half Min = Half Minimum imputation; Mean = Mean imputation; KNN = k-nearest neighbor imputation; LLS = local least squares regression imputation; BPCA = Bayesian principal components analysis imputation; SVD = singular value decomposition imputation; RF = random forest imputation). Each dot represents a correlation coefficient between plasma and serum for one glycan. The diagonal line in each plot depicts equal correlation in the full and imputed data sets. The dashed lines show correlation of 0. Filled dots indicate false positives of a MANOVA using the imputed data. False positives were defined as glycans found to differ significantly between Cases and Controls with the imputed data set but not with the full data set.

each consisting of 100 compounds for each combination of mean differences, correlation level and missingness level. Missing values were imputed for each data set of 100 compounds. A MANOVA was applied to log-transformed data and the average numbers of true and false positives and true and false negatives were calculated across the 1000 data sets for each imputation method versus the full data set. Significance was set at $P < 0.05$ without adjusting for multiple testing. Single matrix *t*-tests using non-missing values were also conducted.

Results

Effect on correlation estimation

All imputation methods impacted the correlation between biological matrices observed in the real data sets, acting to reduce the magnitude of the correlation (Figure 1). The largest changes to the correlations occurred with the largest level of missingness. With 50% missing values, some strongly positively correlated compounds reversed direction to weakly negatively correlated after imputation (Figure 1). The simulations corroborated these results showing all imputation methods to cause a general reduction in the magnitude of the correlation. The degree of change was greater for strongly correlated compounds, and also increased with increasing levels of missingness (Figure 2, Supplementary Figure S2). Notably, even the direction of the correlation could change with imputation. Half Minimum had the worst performance, resulting in the largest displacement from the true correlation distribution (Figure 2 and Supplementary Figure S2). Across all levels of correlation and missingness, BPCA and RF were best at preserving the between-matrix correlation (Figure 2 and Supplementary Figure S2).

Impacts on MANOVA results of real data sets

With the real data sets, the level of missingness and the imputation method impacted the results of MANOVAs. In general, as the level of missingness increased, the number of true positives decreased and false positives increased (Figure 3). Mean imputation often yielded the highest number of true positives but also always had the highest number of false positives, sometimes with large numbers of false positives. Half Minimum imputation tended to result in few significant findings, small numbers of both true and false positives. The remaining methods provided a more favorable balance of true and false positives. No one method was consistently best in terms of maximizing the number of true positives while controlling the number of false positives, and the relative performance of the methods varied depending on the data set and level of missing values. With the Xenograft data set, RF, BPCA and SVD all performed well with high and similar numbers of true positives and low numbers of false positives at all missing value levels. RF continued to have high true positives and low false positives at 10% and 25% missing for the lung cancer serum/plasma glycomics data set (LC-SP) data set but performed relatively poorly at 50% missing. SVD yielded similar performance to RF at 25% missing, but had the fewest true positives at 10% and 50% missing. Interestingly, with the lung cancer blood glycomics data set (LC-DBS) data set, LLS and SVD had the highest true positives at 10%, but LLS had more false positives. At 25% and 50% missing, KNN and BPCA had similar relatively high numbers of true positives, but false positives were better controlled with BPCA.

Finally, we compared the MANOVA results using imputed data to single matrix *t*-tests with a Bonferroni adjustment. For almost all data sets and levels of missingness, the MANOVA procedures using imputed data had higher numbers of true positives than the single matrix tests. False positives were largely comparable between the single matrix tests and MANOVAs except for Mean imputation at 10% and 25% missing.

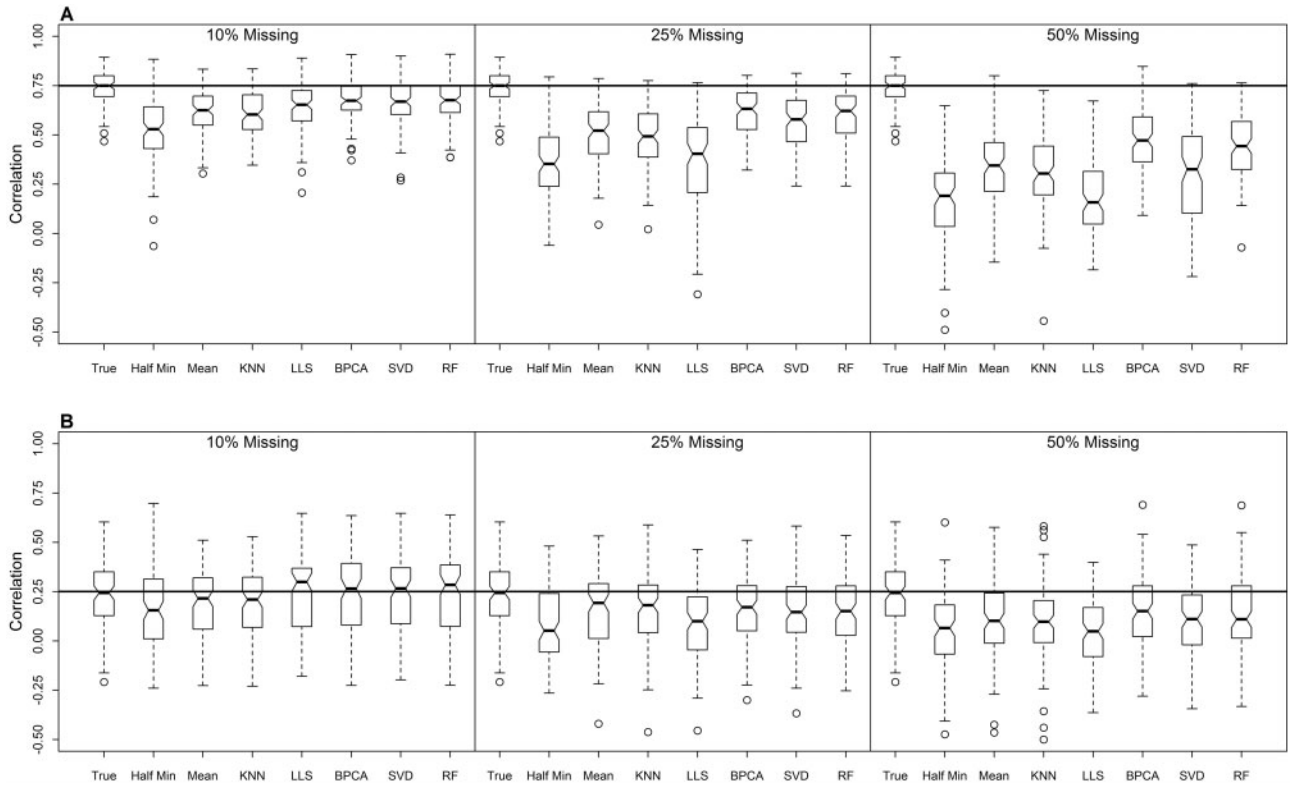

Figure 2: Comparison of correlation between two biological matrices for 100 simulated compounds with 10%, 25% and 50% missing values after imputation using seven methods (Half Min = Half Minimum imputation; Mean = Mean imputation; KNN = *k*-nearest neighbor imputation; LLS = local least squares regression imputation; BPCA = Bayesian principal components analysis imputation; SVD = singular value decomposition imputation; RF = random forest imputation). True shows the distribution of the correlations of the full data set. Data shown here were simulated from a bivariate normal distribution with a mean of 0 and variance of 1 with correlations (rho) indicated by the solid horizontal lines of (**A**) 0.75 and (**B**) 0.25 and then exponentiated.

At 50% missing, the single matrix tests had fewer false positives for the Xenograft and LC-DBS data sets but more for the LC-SP data set than the MANOVA procedures.

True and false negatives mirrored the true and false positive findings (Supplementary Figure S3). Half Minimum, which had the lowest power of the methods, tended to have large numbers of true and false negatives, while Mean imputation, which yielded many rejections, had relatively small numbers of true and false negatives with the remaining methods intermediate to these two methods.

Simulation study of impacts to MANOVA results

Impact of level of missingness

The simulation study provided a better understanding of the effects of the level of missingness and differences among the imputation methods on MANOVA results. There were several notable patterns. First, consistent with the real data set findings, for all simulation scenarios, the number of false positives generally increased for all imputation methods as the level of missingness increased (Figures 4 and 5). BPCA and Half Minimum were the exceptions, maintaining a consistent false positive rate regardless of the level of missing values. As with the real data sets, Mean imputation showed a substantial increase in false positives with increasing missingness, and Half Minimum had little power to detect significant compounds.

At 10% missing values, the remaining imputation methods were broadly similar in terms of the number of true and false positives. SVD, however, did identify a few more true positives while maintaining the smallest number of false positives. Differences among the imputation methods became more evident at 25% and 50% missing. At 25% missing, KNN, SVD and RF had relatively large numbers of true positives, but among these three, RF had the smallest number of false positives. LLS detected the smallest number of true positives other than Half Minimum at 25%, but at 50% missing, BPCA had fewer true positives. At 50% missing, KNN continued to yield a high number of true positives but with a fair number of false positives. SVD and LLS had similar numbers of false positives as KNN at 50% missing but fewer true positives. BPCA had low power at 50% with small numbers of true and false positives. At 50% missing, RF was intermediate, detecting a moderate number of true positives while controlling the number of false positives (Figures 4 and 5). These patterns were consistent regardless of whether Cases differed from Controls in both or only one of the simulated biological matrices. The true and false negative rates mirrored the true and false positives and are shown in Supplementary Figures S4 and S5.

Impact of differential effect sizes

The effect of imputing missing values on MANOVA results differed depending on whether Cases and Controls had different means in both biological matrices versus in only one of the matrices. For the full data sets, when both biological matrices differed in the same direction (i.e. Cases had higher means than Controls in both biological matrices), the largest number of significant results occurred when the matrices were strongly negatively correlated (Figure 4).

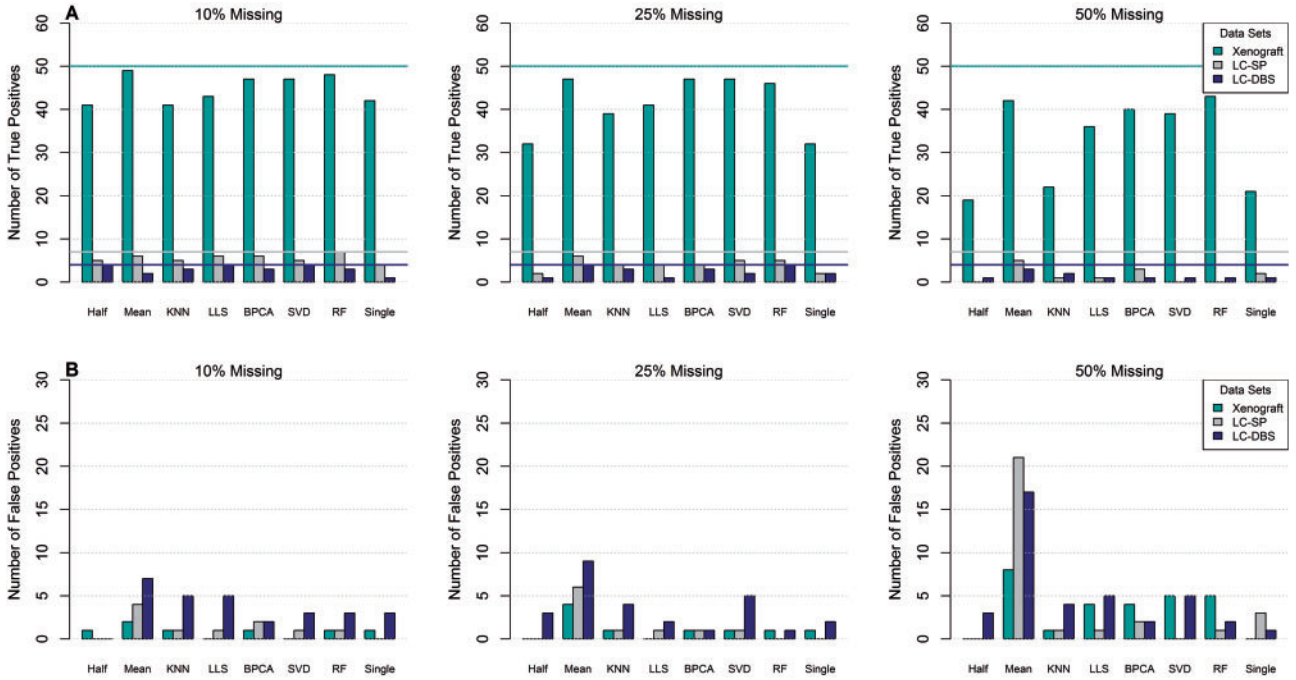

Figure 3: Number of true positives and false positives for MANOVAs applied to three real omics data sets with 10%, 25% and 50% induced missing values imputed using seven methods (Half = Half Minimum imputation; Mean = Mean imputation; KNN = k-nearest neighbor imputation; LLS = local least squares regression imputation; BPCA = Bayesian principal components analysis imputation; SVD = singular value decomposition imputation; RF = random forest imputation). Single shows single matrix results of *t*-tests using non-missing data with a Bonferroni adjustment to account for testing multiple matrices. The thick, horizontal lines indicate the number of significant ($P < 0.05$) MANOVA tests based on the full data set. Xenograft = renal cell carcinoma xenograft metabolomics data set; LC-SP = lung cancer serum/plasma glycomics data set; LC-DBS = lung cancer blood glycomics data set. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

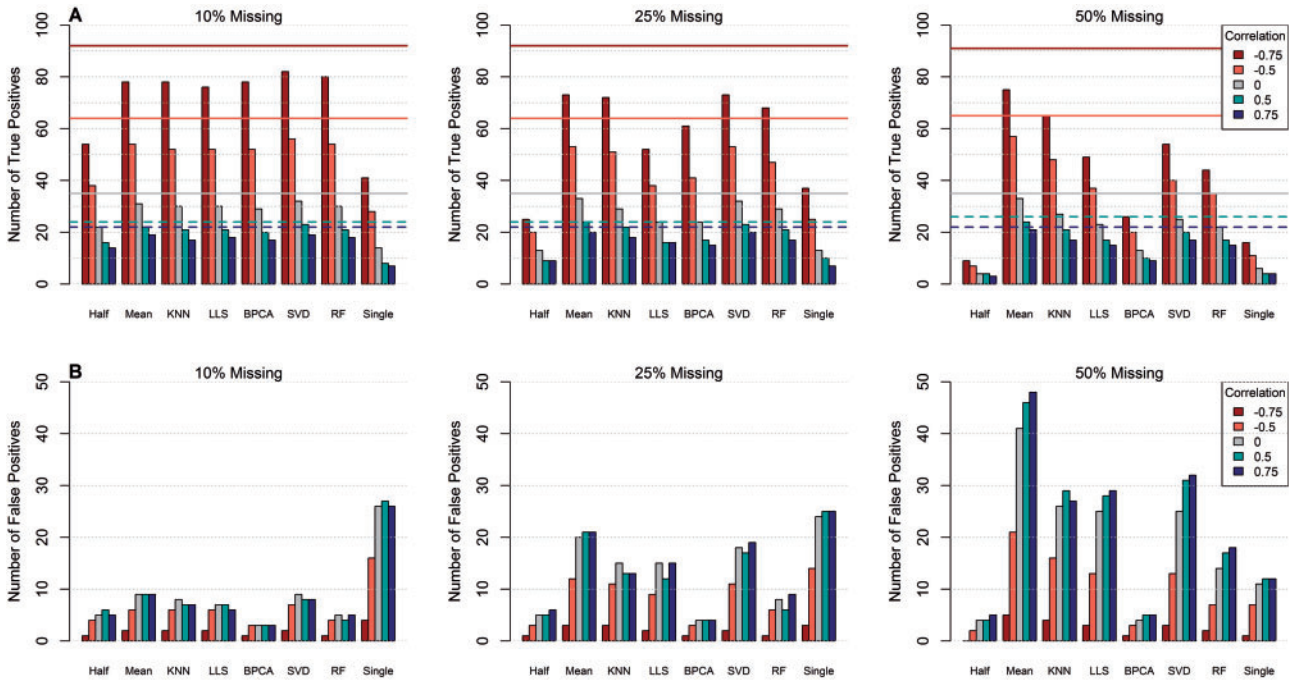

Figure 4: MANOVA results for simulated data sets in which Cases differ from Controls in both biological matrices. The number of (**A**) true positives and (**B**) false positives for MANOVAs applied to simulated data sets with 10%, 25% and 50% induced missing values imputed using seven methods (Half Min = Half Minimum imputation; Mean = Mean imputation; KNN = k-nearest neighbor imputation; LLS = local least squares regression imputation; BPCA = Bayesian principal components analysis imputation; SVD = singular value decomposition imputation; RF = random forest imputation) are shown. 'Single' shows single matrix results of *t*-tests using non-missing data with a Bonferroni adjustment to account for testing multiple matrices. The thick dashed or solid horizontal lines indicate the number of significant ($P < 0.05$) MANOVA tests based on the full data set; the line colors correspond to the correlations in the legend. Data were simulated from a bivariate normal distribution with a variance of 1 and between matrix correlations of ± 0.75, ± 0.50 and 0 and then exponentiated. The mean of the Controls in each biological matrix was 0. The mean of the Cases was set at 0.5 in both biological matrices. For each set of parameters, 1000 data sets, each consisting of 100 compounds, were simulated. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

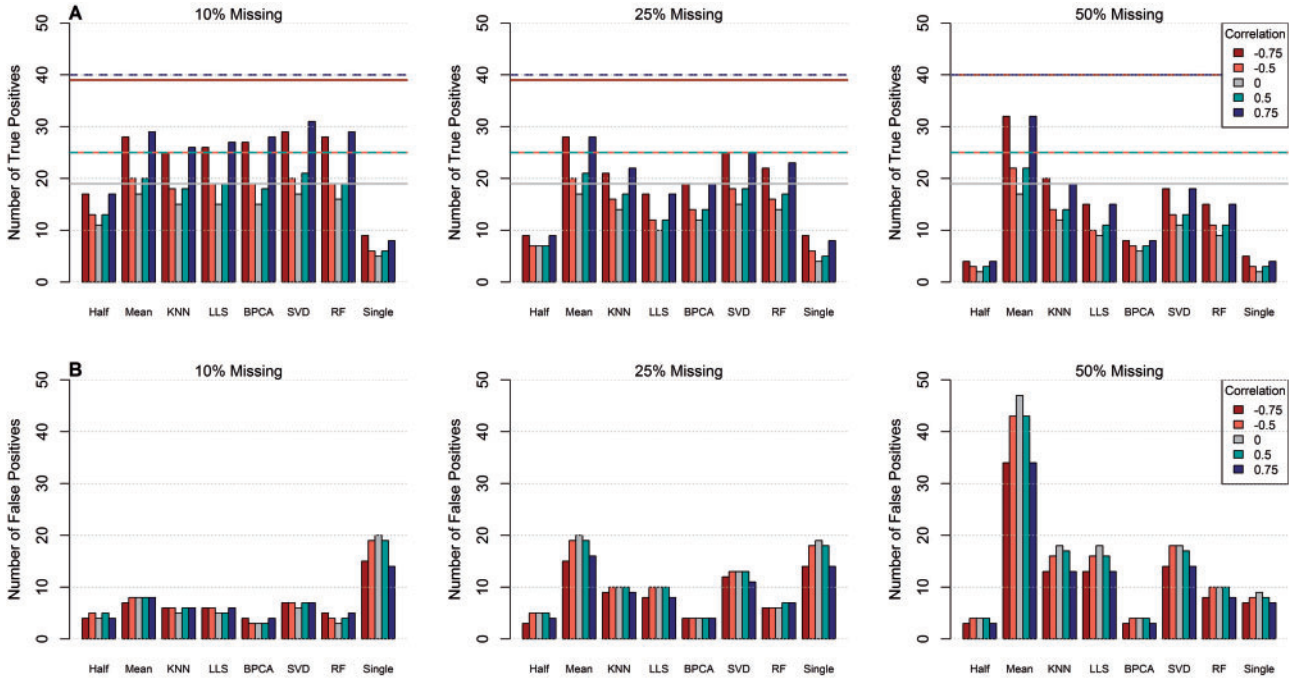

Figure 5: MANOVA results for simulated data sets in which Cases differ from Controls in only one biological matrix. The number of (**A**) true positives and (**B**) false positives for MANOVAs applied to simulated data sets with 10%, 25% and 50% induced missing values imputed using seven methods (Half = Half Minimum imputation; Mean = Mean imputation; KNN = k-nearest neighbor imputation; LLS = local least squares regression imputation; BPCA = Bayesian principal components analysis imputation; SVD = singular value decomposition imputation; RF = random forest imputation) are shown. The thick dashed or solid horizontal lines indicate the number of significant ($P < 0.05$) MANOVA tests based on the full data set; the line colors correspond to the correlations in the legend. Single shows single matrix results of *t*-tests using non-missing data with a Bonferroni adjustment to account for testing multiple matrices. Data were simulated from a bivariate normal distribution with a variance of 1 and between-matrix correlations of ± 0.75, ± 0.50 and 0 and then exponentiated. The mean of the Controls in each biological matrix was 0. The mean of the Cases was set at 0.5 in one biological matrix. For each set of parameters, 1000 data sets each consisting of 100 compounds were simulated. The thick, colored horizontal lines in the true positive figures (**A**) show the numbers of significant ($P < 0.05$) MANOVAs based on the full data sets. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

As the correlation increases from strongly negative to strongly positive, the number of significant results declined (Figure 4). This pattern was also apparent in the imputed data sets. However, with the imputed data sets there were large numbers of false positives for large amounts of missing data, often in equal or greater numbers to the true positives (Figure 4). This effect was largest for strongly positively correlated matrices. Further, the reduction in the number of true positives with increasing missingness was greatest for strongly negatively correlated matrices (Figure 4).

The dynamics were different when Cases differed from Controls in only one of the biological matrices. First, with the full data set, the number of significant results increased as the between-matrix correlation became stronger either positively or negatively (Figure 5). False positives were highest for moderately correlated matrices. Unlike the results when Cases differed from Controls in both matrices, increasing missingness reduced the number of true positives by a similar amount regardless of the correlation.

Comparison of real data and simulation study results

Interpreting the real data analysis results in the context of the simulation results is complicated by (1) not knowing the true differences between Cases and Controls for the real data sets, (2) the varied correlation levels and direction among the biological matrices and (3) the absence of a within-matrix correlation structure for the simulated data that could influence performance of some of the imputation methods examined. Nevertheless, there are some similarities. Increased missing values decreased the number of true positives and increased the number of false positives for both the real and simulated data sets. Half Minimum generally had low numbers of true and false positives while Mean imputation had high numbers of true and false positives. The largest differences between the simulations and real data results were for KNN, which had some of the highest numbers of true positives in the simulated data but was among the lowest for the real data. SVD had fewer false positives with the real data, and BPCA had relatively high numbers of true positives with the real data, with 50% missing but low true positives for the simulated data at this level of missingness. RF and LLS had comparable relative performance in the simulations and real data.

In these analyses, we focused on MANOVA performance. Typically though, investigators independently analyze each biological matrix. Although a comprehensive evaluation of the power of MANOVA versus single matrix analysis methods is outside the scope of this investigation, we did compare MANOVA results to single matrix analyses of the real data sets. For the real data sets with no missing values, the MANOVA detected more differentially regulated compounds for the Xenograft (50 versus 47) and LC_SP data sets (7 versus 4). For LC_DBS, the MANOVA identified one fewer significant compounds than the five identified through individual-matrix analysis. These results indicate that when there are no missing values, MANOVAs can yield higher power than separate analyses of each matrix. When missing values are present, MANOVAs cannot be used unless the missing values are imputed. Imputation can influence the within-subject between-

matrix correlation, leading to potentially undesirable effects on MANOVA results; thus, independent analysis of biological matrices using only observed values could be more favorable. In our simulations, in addition to using MANOVA with imputing missing values, we applied *t*-tests using non-missing values in each of the simulated biological matrices. At 10% and 25% missing, the single matrix procedures had more false positives and fewer true positives than the MANOVAs. At 50%, the single matrix analyses had fewer false positives than MANOVA's on data imputed using Mean, KNN, LSS, SVD and RF but detected fewer true positives (Figures 4 and 5). The smaller sample sizes associated with only using observed values reduced statistical power of the individual matrix analyses (Figures 4 and 5).

Discussion

Imputation is a common approach to handling missing values in MS data, but before using such a strategy, it is important to understand the potential downstream effects of the imputation on analytical results. The effect of various imputation approaches on differential and multivariate analysis results has been extensively studied when analyzing a single biological matrix [7] but has not been evaluated for multivariate analysis across multiple biological matrices. In this study, we investigated the effects of several imputation methods in a multiple biological matrix setting and found that imputation did not preserve the between-biological matrix correlation structure of the true data and substantially impacted significance testing results using MANOVA.

We found that imputing missing values reduced the magnitude of the correlation between biological matrices, resulting in more weakly correlated compounds. In extreme cases with large amounts of missing values, strongly positively correlated compounds could become weakly negatively correlated after imputation. The failure of the imputation to preserve the correlation structure could result from a discrepancy between the assumptions of the imputation method and inherent mechanisms producing missing values in MS-generated omics data. In MS studies, missing values arise from multiple processes including missing at random and missing not at random processes [11]. Half Minimum imputation assumes missing data arises from censoring below a predetermined level of detection set by a signal-to-noise ratio, while the other imputation methods assume that missing data arise at random. Because missing values originate from multiple processes in MS studies, none of these imputation methods fully reflects the genuine missing data mechanisms at work here. This discrepancy could contribute to the failure of the imputation methods to maintain the true correlation structure.

The correlation between variables can substantially affect MANOVA results. For two variables, Cole *et al.* [18] found that when the effect size and direction was similar in the two variables, MANOVA power is highest when the two variables are strongly correlated in the opposite direction (i.e. for positive effects in both variables, power is highest for strongly negatively correlated variables). If, however, one of the variables has a weak or no effect, highest power occurs at strong correlations irrespective of the direction of the correlation. Clearly then, changes in the correlation structure owing to imputation can alter the results of significance testing. All other aspects being the same (i.e. means and variances), reducing the magnitude of the correlation would lead to fewer rejections of no difference when Cases truly differ from Controls in only one of two biological matrices. When there are differences in multiple biological matrices, reducing the magnitude of the correlation would lead to fewer rejections of no difference for strongly negatively correlated matrices and more rejections for strongly positively correlated matrices, assuming mean differences in the same direction in both matrices.

When MS data are available from multiple biological matrices from the same patient, using a testing procedure that draws on information in all biological matrices could be advantageous. However, our results highlight the need to carefully consider the extent and pattern of missingness and the potential impacts of imputation on multivariate analysis results. We found that as the amount of missing data increased, the number of false positives increased and the number of true positives decreased substantially, which is an important issue for high-dimensional data analysis. Similar to other imputation investigations, no one method was universally optimal [7, 11, 28]. The two substitution methods (Half Minimum and Mean) performed poorly. With Half Minimum, any missing values for a compound are imputed with the same small value based on detected compounds, which tends to suppress differences between experimental groups resulting in less statistical power. In contrast, Mean imputation in which missing values are imputed using the mean of the experimental group associated with the missing observation will tend to increase differences between experimental groups as well as reduce the variance, leading to more rejections of the null hypothesis of no difference. The global structure (BPCA and SVD) and local similarity methods (KNN, LLS and RF) were more effective than the substitution methods. By integrating information from multiple similar samples and compounds, these methods better preserved the between-matrix correlation structure and MANOVA properties. Our results underscore the need for further methodological research in multiple areas including developing imputation methods appropriate for multiple biological matrix MS-based omics data with unique missing mechanisms as well as development of analytical methods for hypothesis testing of multivariate data such as Expectation-Maximization approaches or two-part statistics that can accommodate missing values without the need for imputation.

Key Points

- The imputation methods investigated here do not retain the between-biological matrix correlation structure, and the degree of deviation from the true correlation structure increases with increasing missingness.
- Results of MANOVAs are strongly affected by the level of missingness and the imputation method. In particular, the number of true positives decreased and the number of false positives increased as the percentage of missing values increased.
- No single imputation method was universally the best, but the simple substitution methods (Half Minimum and Mean) consistently performed poorly.
- Overall, our study highlights the need to carefully take into account missingness behaviors when an analysis method is chosen.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by the National Institutes of Health (grant P01 AG025532 to K.K.), the National Center for Advancing Translational Sciences, National Institutes of Health (UL1 TR000002) and the UC Davis MIND Institute Intellectual and Developmental Disabilities Research Center, National Institutes of Health (U54 HD079125) and Tobacco Related Disease Research Program (20PT0034).

References

1. Jordan KW, Adkins CB, Su L, *et al*. Comparison of squamous cell carcinoma and adenocarcinoma of the lung by metabolomic analysis of tissue-serum pairs. *Lung Cancer* 2010;**68**: 44–50.
2. Chen YJ, Wang XH, Huang ZZ, *et al*. A study of human bladder cancer by serum and urine metabonomics. *Chin J Anal Chem* 2012;**40**:1322–8.
3. Yonezawa K, Nishiumii S, Kitamoto-Matsuda J, *et al*. Serum and tissue metabolomics of head and neck cancer. *Cancer Genomics Proteomics* 2013;**10**:233–8.
4. Ganti S, Taylor SL, Abu Aboud O, *et al*. Kidney tumor biomarkers revealed by simultaneous multiple matrix metabolomics analysis. *Cancer Res* 2012;**72**:3471–9.
5. Austdal M, Skrastad RB, Gundersen AS, *et al*. Metabolomic biomarkers in serum and urine in women with preeclampsia. *PLoS One* 2014;**9**:e91923
6. Witowski N, Lusczek E, Determan C, Jr, *et al*. A four-compartment metabolomics analysis of the liver, muscle, serum, and urine response to polytrauma with hemorrhagic shock following carbohydrate prefeed. *PLoS One* 2015;**10**:e0124467
7. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics* 2011;**8**:161–74.
8. De Livera AM, Olshansky M, Speed TP. Statistical analysis of metabolomics data. *Methods Mol Biol* 2013;**1055**:291–307.
9. Taylor SL, Leiserowitz GS, Kim K. Accounting for undetected compounds in statistical analyses of mass spectrometry 'omic studies. *Stat Appl Genet Mol Biol* 2013;**12**:703–22.
10. Troyanskaya O, Cantor M, Sherlock G, *et al*. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001;**17**:520–5.
11. Webb-Robertson BJ, Wiberg HK, Matzke MM, *et al*. Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J Proteome Res* 2015;**14**:1993–2001.
12. de Brevern AG, Hazout S, Malpertuy A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics* 2004;**5**:114
13. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Brief Bioinform* 2010;**11**:253–64.
14. Scheel I, Aldrin M, Glad IK, *et al*. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 2005;**21**:4272–9.
15. Gromski PS, Xu Y, Kotze HL, *et al*. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites* 2014;**4**:433–52.
16. Karpievitch Y, Stanley J, Taverner T, *et al*. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* 2009;**25**:2028–34.
17. Webb-Robertson BJ, Matzke MM, Metz TO, *et al*. Sequential projection pursuit principal component analysis–dealing with missing data associated with new -omics technologies. *Biotechniques* 2013;**54**:165–8.
18. Cole DA, Maxwell SE, Arvey R, *et al*. How the power of MANOVA can both increase and deecrease as a function of the intercorrelations among the dependent variables. *Psychol Bull* 1994;**115**:465–74.
19. Ruhaak LR, Taylor SL, Miyamoto S, *et al*. Chip-based nLC-TOF-MS is a highly stable technology for large-scale high-throughput analyses. *Anal Bioanal Chem* 2013;**405**:4953–8.
20. Kim K, Ruhaak LR, Nguyen UT, *et al*. Evaluation of glycomic profiling as a diagnostic biomarker for epithelial ovarian cancer. *Cancer Epidemiol Biomarkers Prev* 2014;**23**:611–21.
21. Ruhaak LR, Miyamoto S, Kelly K, *et al*. N-Glycan profiling of dried blood spots. *Anal Chem* 2012;**84**:396–402.
22. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 2005;**21**:187–98.
23. Oba S, Sato Ma, Takemasa I, *et al*. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* 2003;**19**:2088–96.
24. Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012;**28**:112–18.
25. Hastie T, Tibshirani, R, Narasimhan, B, *et al*. impute: Imputation for microarray data. R Package version 1.34.0. 2013. Available at https://bioconductor.org/packages/release/bioc/html/impute.html.
26. Stacklies W, Redestig H, Scholz M, *et al*. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 2007;**23**:1164–7.
27. Stekhoven DJ. missForest: Nonparametric missing value imputation using random forest. R Package version 1.4. 2013. Available at https://cran.r-project.org/web/packages/missForest/.
28. Brock GN, Shaffer JR, Blakesley RE, *et al*. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinf* 2008;**9**:12.