EFFECTS OF MISSING VALUE IMPUTATION ON DOWN-STREAM ANALYSES IN MICROARRAY DATA

by

Sunghee Oh

BSc, Cheju National University, Republic of Korea, 2001

MA, Yonsei University, Republic of Korea, 2003

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2009

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Sunghee Oh

It was defended on

June 19, 2009

and approved by

Dissertation Advisor:
George C.Tseng, Sc.D
Assistant Professor
Biostatistics and Human Genetics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Jonghyeon Jeong, Ph.D
Associate Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Lan Kong, Ph.D
Assistant Professor
Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
Yan Lin, Ph.D
Assistant Professor
Biostatistics and Human Genetics
Graduate School of Public Health
University of Pittsburgh

EFFECTS OF MISSING VALUE IMPUTATION ON DOWN-STREAM ANALYSES IN MICROARRAY DATA

Sunghee OH, PhD

University of Pittsburgh, 2009

Amongst the high-throughput technologies, DNA microarray experiments provide enormous quantity of genes and arrays with biological information to disease. The studies of gene expression values in various conditions and various organisms in public health have led to the identification of genes to the comparison between tumor and normal, clinically relevant subtypes of tumor, and prognostic signatures and have ultimately provided the potential targets for specific therapy of public health disease. Despite such advances and the popular usage of microarray, the microarray experiments frequently produce multiple missing values due to many flaw factors such as dust, scratches on the slides, insufficient resolution, or hybridization errors on the chips. Thus, gene expression data contains missing entries and a large number of genes may be affected. Unfortunately, many downstream algorithms for gene expression analysis require a complete matrix as an input. Therefore effective missing value imputation methods are needed and have been developed in the literature so far. There exists no uniformly superior imputation method and the performance depends on the structure and nature of a data set. In addition, imputation methods have been mostly compared in terms of variants of RMSEs (Root Mean Squared Error) to compare similarity between true expression values and imputed expression values. The drawback of RMSE-based evaluation is that the measure does not reflect the true biological effect in down-stream analyses.

In this dissertation, we will investigate how missing value imputation process affects the biological result of differentially expressed genes discovery, clustering and classification. Multiple statistical methods in each of the downstream analysis will be considered. Quantitative measures reflecting the true biological effects in each down-stream analysis will be used to evaluate imputation methods and be compared to RMSE-based evaluation.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENT

The work presented in this dissertation would have not been possible without the help of many people. I would like to express an eternal gratitude to several individuals who contributed their support and guidance to complete successfully. First of all, I would like to thank my advisor and committee chair, Dr.Tseng. During the PhD degree, I learned much statistical analyses and research methods about Biostatistics and Bioinformatics on working in his group and I am convinced that this knowledge will help me in the future. He always pushed me to achieve as much as I could, and taught me how to become a better biostatistician. He devoted his time to help me in navigating the dissertation. His ideas and tremendous support had a major influence on this dissertation. I would like to all committee members, Dr.Jeong, Dr.Lin, and Dr.Kong for providing their expertise, valuable advice, time and reviewing my dissertation during busy semester schedule. I would like to express thanks a bunch to Dr.Kong for sharing her valuable statistical advice, Dr.Jeong for mentoring me during the entire PhD degree, Dr.Lin for teaching me how to think creatively and logically on proposal and defense. I learned a great deal from all committee members. Thank you again.

To Dr.Sibille, I sincerely appreciate all support during PhD studies and his research field, which provided me to have had an invaluable opportunity analyzing depression and aging mouse and human data. It was a great training to play a crucial role as a stepping stone to the current

research. Thank you, Dr.Sibille. To Dr.Brock and Don Kang, I would like to thank for their assistance and valuable comments for this MV project.

I have heartfelt appreciation for all my friends, group lab members and Dr.Sibille's lab members for being a good colleague on research and for always being beside me and offering their belief in me at times when I doubted myself.

I would like to especially thank my parents. None of this would have been possible without my loving parents and family. I share this accomplishment with them.

Finally, I dedicate this dissertation to my lord. Thank you my lord.

1.0 INTRODUCTION

1.1 THE BACKGROUND OF MICROARRAY EXPRESSION DATA

The microarray experiment is a new technology to investigate the expression levels of thousands of genes simultaneously. It has become one of the most indispensable tools that many biologists use to monitor genome wide expression levels of genes in a given organism. A microarray is typically a glass slide, which contains thousands of spots and each spot may contain a few million copies of identical DNA molecules that uniquely correspond to a gene. There are mainly two basic microarray technologies, cDNA array of dual-channel by Stanford group and high-density oligonucleotide arrays of one channel pioneered by Affymetrix. Each technology has its own merits and demerits. Figure 1 represents the schematic illustration of how microarrays experiments are performed for cDNA (a) and Oligonucleotide (GeneChip experiment) (b) microarrays. The main difference between two different platforms is how the genes are represented on the arrays, and the way that relative abundances of the transcripts are calculated. The cDNA microarray has two-color hybridization to be able to eliminate array to array noise and it is cheaper than Affymetrix. While cDNA produces relative abundant levels for target mRNAs corresponding to all probes on the array, the oligonucleotide array has only one sample, which hybridizes to a single array rather than target and reference sample and then returns an absolute mRNA level. In both platforms, the microarray gene expression data are represented as a G (gene) × S (sample) matrix, $D = \{D_{gs}\}$,where the entry is the mRNA expression level of gene g in experiment (sample) s.

1.2 THE NEED OF MISSING VALUE IMPUTATIONS ON MICROARRAY EXPRESSION DATA

In the microarray experiment as mentioned in the previous subsection, missing values (MVs) frequently occur from various sources, such as dust, scratches on slide, insufficient resolution, or hybridization error, and etc. However, most statistical downstream analyses for microarray data such as DE (differentially expressed) gene detection, clustering and classification analyses require a complete data as the input. A naïve solution is simply to ignore and delete genes with missing values as a pre-processing step or replace missing entries with zero value, however, it may result in the loss of much critical information as it is known that microarray datasets have usually more than 5% missing values and up to 90% genes might be affected. To utilize the information of genes with missing values, missing entries can be substituted with estimated values using more robust imputation method. Even though numerous imputation methods have been introduced and are constantly developed during the past decade, no uniformly best method exists.(Bo, et al., 2004; Brock, et al., 2008; Kim, et al., 2006; Oba, et al., 2003; Troyanskaya, et al., 2001)The main reason is when using imputation methods for missing values that the performance of an imputation method strongly depends on the structure, nature and complexity of the data.

1.3 MV IMPUTATION METHODS AND PERFORMANCE MEASURES

Imputation methods have been usually evaluated by performance measures, such as RMSE. In such an evaluation, a complete data matrix with no missing value is used as a reference, a certain percentage of missing values are randomly generated and the missing values are imputed by a given MV imputation method. Variants of RMSE are used in the literature to quantify the differences of imputed values and the original true values. In an early study, (Troyanskaya, et al., 2001) examined a comparative study of three imputation methods (KNN-impute, SVD-impute, Row-avg) for the estimation of missing values in gene microarray data with respect to the normalized RMSE by dividing it by average value over all observations in the true complete dataset. This measure is denoted here by NRMSE. (See method)He suggested KNN-Impute is a robust imputation method when comparing to Row-avg and SVD-impute. Even though KNN-Impute has been used as a popular imputation method due to its simplicity and fast computation so far, some recent papers have proposed further improved imputation methods. (Nguyen et al., 2004) suggested regression methods, which multiple imputations via ordinary least squares (OLS) and missing value prediction using partial least squares which accuracy for PLS imputation is higher for some ranges beyond moderate expression. From the results, they showed that KNN-impute is still more accurate, compared to PLS regression methods, when the true expression is near the mean, however outside of this range, PLS outperforms KNN-impute. (Oba, et al., 2003)suggested an imputation method based on Bayesian principal component analysis (BPCA). In the paper, estimation ability of BPCA is overall the best among KNN-impute, SVM-impute, and BPCA method. They used a normalized RMSE by dividing it by the standard deviation of the values corresponding to missing entries in the true complete dataset. We call it NRMSE2. (See method.) However, since BPCA assumes only a global covariance structure, the estimation with BPCA may not be accurate if genes have dominant local similarity structures and KNN-impute will be suitable in the case. (Ouyang, et al., 2004) proposed GMC-impute based on GMC (Gaussian Mixture Clustering) and model averaging. When using GMC-impute, the microarray data are assumed being generated by a Gaussian mixture of some number of components. GMC-impute shows better performance compared to naïve methods such as ZERO-impute, Row-avg, Col-avg, KNN-impute and SVM-impute in terms of normalized RMSE by dividing it by the root mean square of all the observations corresponding to missing entries in the true complete dataset in the study. We call it NRMSE3. (See method) (Bo, et al., 2004)introduced LSA-impute which is based on the least squares principle and utilizes correlations between both genes and arrays. The accuracy of LSA-impute is higher than that of KNN-impute in terms of RMSD (Root Mean Squared Deviation) between the true values and the estimated values to show how well predicted the missing values are. This measure is not a normalized measure. (Kim, et al., 2006)suggested a local least squares imputation method (LLS-impute) to represent a target gene that has missing values as a linear combination of similar genes. The proposed LLS-impute method shows competitive results compared to KNN-impute and BPCA using NRMSE. This normalized RMSE is the same as (Oba, et al., 2003).

Novel methods for missing value imputation have been constantly developed during the last decade. Moreover, most of previous works on MV have been evaluated a MV method in terms of RMSE (Root Mean Squared Error), when comparing the new methods to the existing methods. Recently, a few papers have started to seriously take into account the influence of imputation methods on downstream analyses (de Brevern, et al., 2004; Jornsten, et al., 2005; Scheel, et al., 2005; Tuikkala, et al., 2006; Wang, et al., 2006). The results are, however, neither comprehensive nor conclusive. In other words, they employed the small number of datasets and methods. Additionally, a study is focused on a down-stream analysis. Even thougth a few people evoked an importance of biological impact on MV Imputation more recently, the take-home message still remained. Hence, in this dissertation, I will examine the biological impact assessments as well as classical RMSE-based measurements in comparison study of missing imputation methods to serve an insightful framework on MV imputation study. The detailed descriptions of MV methods and performance measures will be explained in the section 2.

1.4 MOTIVATION AND RESEARCH DESIGN OF OUR COMPARATIVE STUDY

Most evaluation of missing value imputation methods has been addressed by RMSE-based measures, instead of considering true biological impacts on down-stream analyses. Recently, a few people have issued this topic that the best imputation method detected by classical evaluation such as RMSE-based do not guarantee the smallest error to the impact of various statistical downstream analyses such as

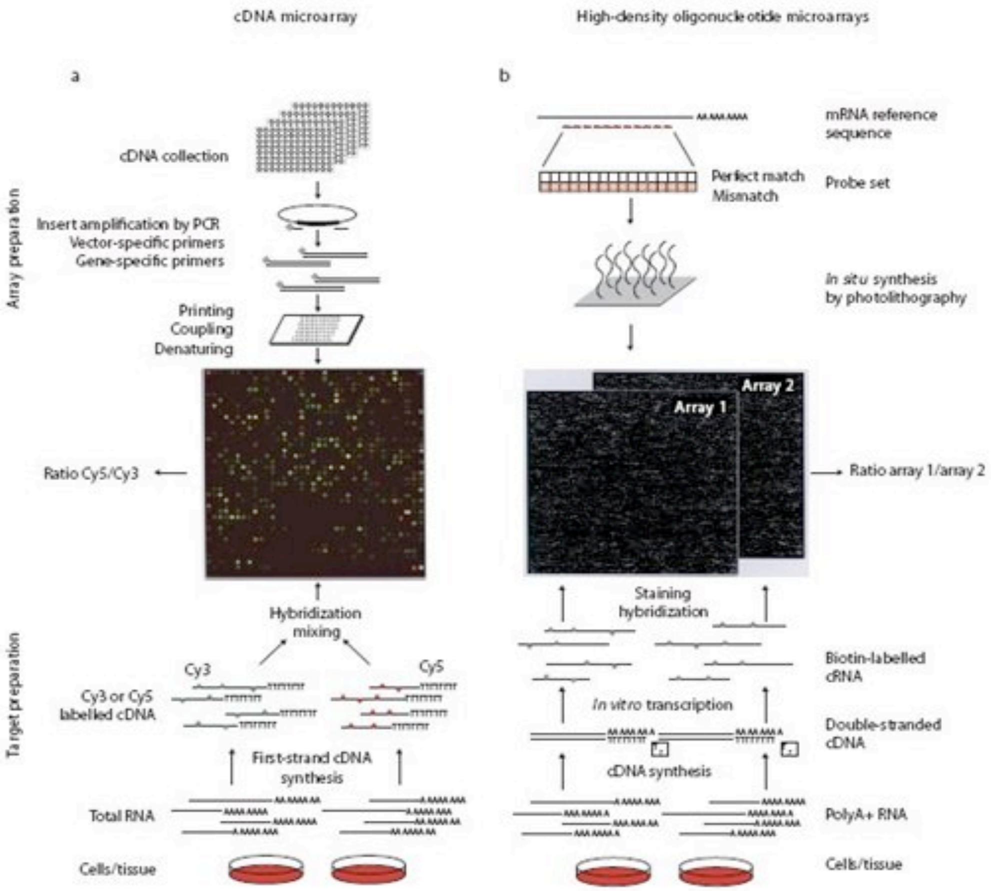


Figure 1.**cDNA and Oligonucleotide microarray.**It represents how experiments are performed for cDNA (left) and Oligonucleotide (right) microarrays. In the top, it shows how microarrays are manufactured; and in the bottom, how RNA samples are obtained. In the middle, we can see images obtained after RNA samples hybridize to the microarrays. For cDNA microarrays (left), each dot represents a probe, and the red (or green) colors are proportional to the counts of RNA hybridized to that probe in the reference (or control) samples. Similarly, the intensity of white dots in Oligonucleotide arrays (right) represents the counts of RNA hybridized to that probe. Figure reproduced from (Simon, et al., 2004).

most common differentially expressed (DE) gene detection(Jornsten, et al., 2005).(Scheel, et al., 2005)examined the impact of imputation on the detection of DE (differentially expressed) genes using SAM and ANOVA. They proposed a novel imputation method, linear model based imputation (LinImp) as well as to have compared existing methods with KNN-impute and LSA-impute. This comparative investigation covers false negative gene list as a biological impact measure. Specifically they counted the number of genes to be falsely declared as non-significant genes compared to gold-standard gene list, where gold standard gene list is defined by the significantly differentially expressed gene list in complete dataset. The focus has been only on differentially expressed gene detection in this paper. However, to draw a more general conclusion to effects of missing values on down-stream analyses, impacts to missing values in classification and other down-stream analyses as well as DE gene detection are needed to present. In clustering analysis, (de Brevern, et al., 2004) investigated the effects of missing value imputation on the stability of gene groups by hierarchical clustering using Conserved Pairs Proportion (CPP). However, in the paper, they presented KNN-imputation method is the most efficient replacement for missing value even though other further sophisticated imputation methods have been studied without ceasing. The limitation of this study is that they only carried out less powerful and inefficient imputation methods such as KNN-impute (KNN.e) and Zero-impute. In classification analysis and functional modules,(Wang, et al., 2006)demonstrated the effects of missing values imputation methods posterior to down-stream analyses. They compared the accuracy rates of three different classifiers, KNN, SVM, and CART classifier on down-stream analyses after imputing missing values using various imputation methods such as ZERO-impute, KNN-impute, LLS-impute, and BPCA in investigating which imputation tool is most robust. (Tuikkala, et al., 2006) investigated the impact of missing value imputation on K-means clustering and interpretation of GO terms from gene expression microarray data. Using 4 imputation methods including naïve method and BPCA, they explored how the agreement of estimated values for missing entries with the original data values and the original clustering results are returned by each imputation method. They employed ADBP (Average distance between partitions of genes) as a biological impact measure as well as normalized RMSE as a classical performance measure. 4 MV imputation methods including average methods and one clustering method are not sufficient to draw a conclusion in comparative study of MV methods.

More recently, (Brock, et al., 2008)showed a more comprehensive comparative study and proposed two selection schemes for selecting the best MV imputation method. Naïve and competitive imputation methods from the previous papers are utilized in this study. Under the assumption that the best MV imputation method depends on the structure and nature of input data, a series of data sets from various experimental designs (two-group comparison, multi-exposure and time series) are analyzed. A log-transformed version of RMSE (named LRMSE) is used as the performance evaluation measure. Through evaluation by LRMSE, they proposed two useful selection schemes of imputation methods, EBS (Entropy based selection) and STS (Simulation based self-training selection). In this comprehensive comparative study, they concluded that there is no universally best MV imputation method although three top methods such as LLS-impute, LSA-impute, and BPCA are very competitive and the accuracy on MV imputation depends on structure and nature of given data set. Even though the study tried to perform a large-scale comparative study, their focus is limited by RMSE-based evaluation. Thus, the previous works on MV imputation methods presents some partial conclusions using smaller number of data sets, comparing fewer MV methods, including fewer down-stream analysis methods in each category or applying inadequate evaluation indices. Hereby, the purpose of this dissertation is to provide a comprehensive comparative analysis to examine the biological impact of MV imputation in all three areas of down-stream analyses by addressing three major aims: (Aim 1A) To investigate whether applying different RMSE measures affects the performance ranking and decision of MV imputation method selection.; (Aim IB) To investigate whether applying different down-stream analysis methods in each category (i.e. SAM, LIMMA and t-test+BH for DE gene detection; LDA, KNN,SVM, and PAM for classification ;K-means, SOM for gene clustering) affects the performance ranking and decision of MV imputation method selection. (Aim 2) If selection of RMSE measure greatly affects the selection of MV imputation method in Aim 1A, investigate which RMSE measure is more consistent (correlated) with the biological impact measure. (Aim 3) Evaluate the consistency and correlation of the best RMSE measure (determined by Aim2) with the biological impact measures in performance ranking and selecting the best MV imputation method. In the beginning, we considered 10 imputation methods including naïve methods of column and row average method. These two methods are obviously of bad performance and are removed. Eight remaining MV imputation methods will keep being discussed in further sections. Eight MV imputation methods ($1 \leq m \leq M=8$; KNN.e, KNN.c, SVD, OLS, PLS, LSA, LLS and BPCA) are considered, eight data sets ($1 \leq d \leq D_1=8$) for DE gene detection and classification and six data sets ($1 \leq d \leq D_2=3$) for gene clustering are evaluated, four missing value percentages ($1 \leq p \leq P=4$; (r_1, r_2, r_3, r_4)=(1%, 5%, 10% and 20%)) are considered and finally 100 independent simulations ($1 \leq n \leq N=100$) are performed. In total, 8×11×4×100=35,200 times of random deletion from complete data matrix and then missing value imputation need to be performed. Due to the already high demand of computing, we skip the procedure of finding the optimal parameter for each MV imputation method in each data set and use the optimal parameters in the comparative study by Brock et al. (2008). For the optimal parameter in KNN and SVM classification, we fixed with the K=5 and linear kernel function, respectively after doing several simulation examinations, from k=1 to k=15 and 4 kernel functions using GOL, ALO, and LUO data set. To investigate quantitative and biological criteria for deciding which MV imputation methods perform better, three RMSE measures (NRMSE, LRMSE and RAE), three DE detection methods (SAM, LIMMA and t-test+BH), four classification methods (LDA, KNN, PAM, SVM) and two gene clustering methods (K-means and SOM) are considered. In gene clustering, the number of clusters is usually not known and usually difficult to estimate from the data. We perform k=5, 10 and 15 to select the best. Therefore, we have 8×11×4×100×3=105,600. RMSE evaluations, 8×8×4×100×3=76,800 DE gene detection evaluations, 8×8×4×100×4=102,400 classification evaluations and 8×3×4×100×2×3=57,600 for gene clustering evaluations. Throughout above proposed research design and 3 main Aims, we believe that the conclusions could provide an insightful conclusion for biological impact of MV imputation on down-stream analyses. To investigate the three Aims above, we apply Spearman's rank correlation to quantify the consistency of selecting (ordering) MV imputation methods given any two criteria of either RMSE measures or biological impact measures. For example, figure 3 shows the plot containing averaged rank of each imputation method for N=100 times using LRMSE and

BLCI. The exact consistency score from Spearman correlation is formulated below. For Aim 1A, we define,

$$\begin{aligned} r_{dpnij}^{RMSE\times RMSE} &= cor_{sp}\Big((RMSE_{1dpni}, RMSE_{2dpni},\ldots, RMSE_{Mdpni}),(RMSE_{1dpnj}, RMSE_{2dpnj},\ldots, RMSE_{Mdpnj})\Big) \\ \tilde{r}_{dpij}^{RMSE\times RMSE} &= \text{median of } \Big\{r_{dpnij}^{RMSE\times RMSE}, 1 \le n \le N\Big\}, \end{aligned}$$

where $RMSE_{mdpni}$ is the RMSE measure from MV imputation method m, data set d, MV percentage p, and simulation n and the RMSE measure i(i=1 meaning NRMSE, 2 for LRMSE and 3 for RAE). The measure $cor_{sp}(\bullet,\bullet)$ is obtained by Spearman's rank correlation. Intuitively, if $\tilde{r}_{dpij}^{RMSE\times RMSE} = 1$, the two RMSE measures i and j give exactly the same rank order of the M=8 MV imputation methods and are considered consistent in MV imputation method selection.

Similarly, we define the consistency score for any two down-stream analysis measures in DE gene detection (BLCI), classification (YI) and clustering (ARI) for Aim 1B as

$$\begin{aligned} r_{dpnij}^{BLCI\times BLCI} &= cor_{sp}\Big((BLCI_{1dpni}, BLCI_{2dpni},\ldots, BLCI_{Mdpni}),(BLCI_{1dpnj}, BLCI_{2dpnj},\ldots, BLCI_{Mdpnj})\Big) \\ \tilde{r}_{dpij}^{BLCI\times BLCI} &= \text{median of } \Big\{r_{dpnij}^{BLCI\times BLCI}, 1 \le n \le N\Big\}, \end{aligned}$$

$$\begin{aligned} r_{dpnij}^{YI\times YI} &= cor_{sp}\Big((YI_{1dpni}, YI_{2dpni},\ldots, YI_{Mdpni}),(YI_{1dpnj}, YI_{2dpnj},\ldots, YI_{Mdpnj})\Big) \\ \tilde{r}_{dpij}^{YI\times YI} &= \text{median of } \Big\{r_{dpnij}^{YI\times YI}, 1 \le n \le N\Big\}, \end{aligned}$$

$$\begin{aligned} r_{dpnij}^{ARI\times ARI} &= cor_{sp}\Big((ARI_{1dpni}, ARI_{2dpni},\ldots, ARI_{Mdpni}),(ARI_{1dpnj}, ARI_{2dpnj},\ldots, ARI_{Mdpnj})\Big) \\ \tilde{r}_{dpij}^{ARI\times ARI} &= \text{median of } \Big\{r_{dpnij}^{ARI\times ARI}, 1 \le n \le N\Big\}, \end{aligned}$$

, where $BLCI_{mdpni,}$, $YI_{mdpni,}$ and $ARI_{mdpni,}$ are the *BLCI*, *YI* or *ARI* measures from MV imputation method m, data set d, MV percentage p, and simulation n and the different selections of down-stream analysis i (SAM, LIMMA, t-test+BH; LDA, KNN, PAM, SVM; K-means, SOM, hierarchical clustering).

Finally, we define the consistency measure for RMSE and biological impact measures as

$$r_{dpnij}^{RMSE\times BLCI} = cor_{sp}\Big((RMSE_{1dpni}, RMSE_{2dpni}, ..., RMSE_{Mdpni}), (BLCI_{1dpnj}, BLCI_{2dpnj}, ..., BLCI_{Mdpnj})\Big)$$

$$\tilde{r}_{dpij}^{RMSE\times BLCI} = \text{median of } \left\{ r_{dpnij}^{RMSE\times BLCI}, 1 \le n \le N \right\},$$

$$r_{dpnij}^{RMSE\times YI} = cor_{sp}\Big((RMSE_{1dpni}, RMSE_{2dpni}, ..., RMSE_{Mdpni}), (YI_{1dpnj}, YI_{2dpnj}, ..., YI_{Mdpnj})\Big)$$

$$\tilde{r}_{dpij}^{RMSE\times YI} = \text{median of } \left\{ r_{dpnij}^{RMSE\times YI}, 1 \le n \le N \right\},$$

$$r_{dpnij}^{RMSE\times ARI} = cor_{sp}\Big((RMSE_{1dpni}, RMSE_{2dpni}, ..., RMSE_{Mdpni}), (ARI_{1dpnj}, ARI_{2dpnj}, ..., ARI_{Mdpnj})\Big)$$

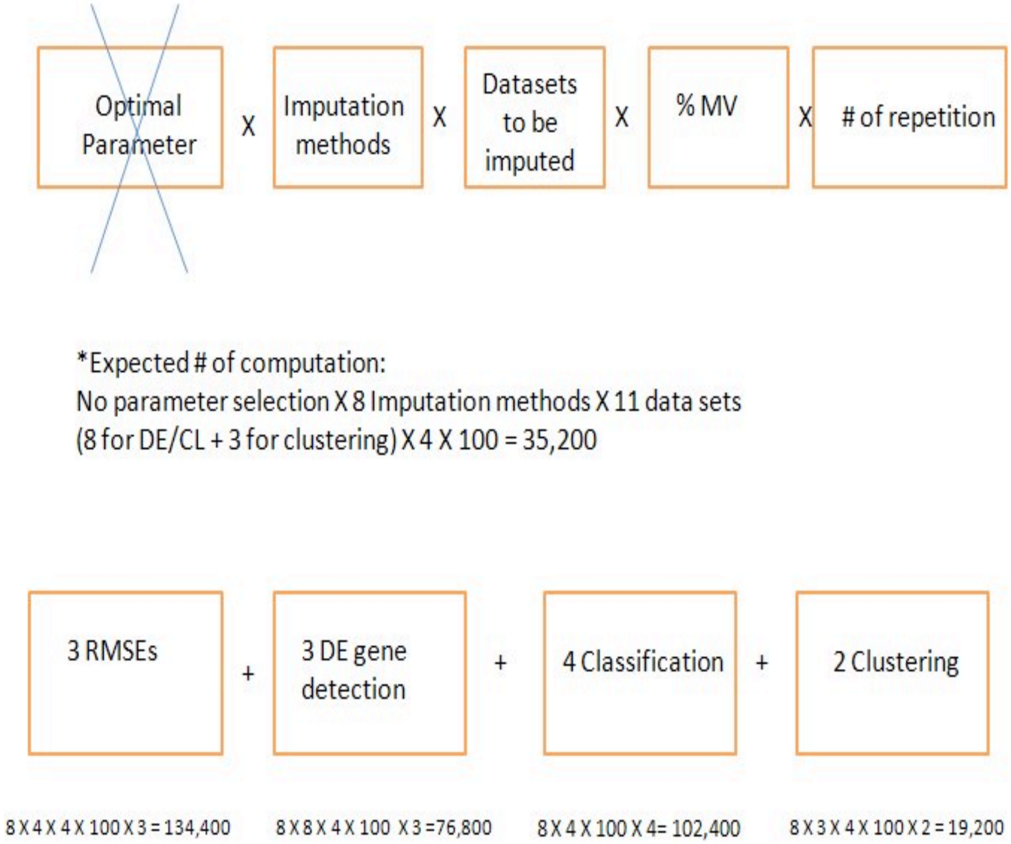
$$\tilde{r}_{dpij}^{RMSE\times ARI} = \text{median of } \left\{ r_{dpnij}^{RMSE\times ARI}, 1 \le n \le N \right\},$$

,where *i* is an index for RMSE measure and *j* for a down-stream analysis method.

In Aim 3, in addition to using the consistency measure for RMSE and biological impact measures for MV imputation method ranking, we further apply the following simple linear regression model to investigate the degree of correlation and the slope:

$$BLCI_{mdpni} = \alpha_{dpnij}^{BLCI} + \beta_{dpnij}^{BLCI} \times RMSE_{mdpnj} + \varepsilon_{mdpnij}$$

,where the linear model interprets how well *RMSE* measures can predict *BLCI* measures and β_{dpnij} reflects the slope in data set *d*, MV proportion *p*, simulation *n*, *RMSE* measure *i* and DE detection method *j*. We further define $\beta_{dpij}^{*(BLCI)} = \text{median of } \left\{ \hat{\beta}_{dpij}^{BLCI}, 1 \le n \le N \right\}$ and the 95% confidence interval $(\beta_{dpij}^{L;(BLCI)}, \beta_{dpij}^{H;(BLCI)})$as the 2.5% and 97.5% quantile of the *N* slope estimate $\left\{ \hat{\beta}_{dpij}^{BLCI}, 1 \le n \le N \right\}$. Intuitively, good missing value imputation results in low RMSE and high BLCI and we expect the β estimates to be negative. When β is negative and large in absolute value, differences of RMSE among different MV imputation methods contribute to real biological impact in BLCI and the method selection by RMSE is meaningful. If β is negative but close to zero, difference of RMSE does not affect the biological impact measure in BLCI and the selection by RMSE is redundant.



*Expected # of computation:
No parameter selection X 8 Imputation methods X 11 data sets
(8 for DE/CL + 3 for clustering) X 4 X 100 = 35,200

Figure 2.**Research Design.** It summarizes our comparative study

2.0 METHODS

2.1 DATA COLLECTION AND PREPROCESSING

For each data set that has MVs in the step to collect dataset, we deleted genes and samples with missing entries so that a complete data matrix without MVs can be used in the study. Thus, original size represents the size of matrix prior to pre-processing to delete missing values, whereas used size is matched by the matrix to be analyzed and evaluated in this comparative study. For clarity, Table 1 lists important features and characteristics of the data sets more details.

Table 1.The description of 11 datasets used in down-stream analyses.

DE/CL	GOL	7129X72	1994X72	AML = 25 ALL = 47	Affy	Min(expr)*=0 Max(expr)*=16.12
DE/CL	ALO	6500X62	2000X62	Colon cancer = 40 Normal = 22	Affy	Min(expr)*= 2.54 Max(expr)*=14.35
DE/CL	LUO	6500X25	6433X25	PA*=16 BP*=9	cDNA	Min(expr)*=0 Max(expr)*=9.6
DE/CL	SIN	12600X102	1662X102	PT*=52 AP*=50	Affy	Min(expr)*=1 Max(expr)*=14.10
DE/CL	LAP	39009 X112	3098X71	PC*=62 M-PC*=9	cDNA	Min(expr)*=-8.83 Max(expr)*=12.36
DE/CL	VAN	25000X117	3196 X 97	BC*=51 BC-M*=46	cDNA	Min(expr)*=0.87 Max(expr)*=1.20
DE/CL	YU	37777X152	2532X89	Prostate cancer=66 Normal=23	Affy	Min(expr)*=0 Max(expr)*=13.04
DE/CL	BEE	7129X96	3577 X 96	OD*=10 AC*=86	Affy	Min(expr)*=-3.32 Max(expr)*=17.01

Table 1 continued.

Gene clustering	SP.AFA	7681X18	4480X18	Time series,cyclic	cDNA	Min(expr)*=-2.71 Max(expr)*=4.76
Gene clustering	SP.ELU	7681X14	5766X14	Time series, cyclic	cDNA	Min(expr)*=-6.22 Max(expr)*=4.95
Gene clustering	CAU	4682X45	4616X45	Multi exposure Time series	cDNA	Min(expr)*=3.00 Max(expr)*=11.44

DE: Differentially expressed gene detection, CL: Classification. PA*=Prostate adenocarcinoma, BP*= benign prostatic hyperplasiaz specimens, PT*=Prostate tumor samples, AP*=Adjacent Prostate, OD*= organ donors normal samples, AC*= adenocarcinoma, BC*=Breast Cancer, BC-M*=Breast Cancer with metastasis, PC*=Prostate Cancer, M-PC*= Metastasis Prostate Cancer in lymph node, Min(expr)*=minimum expression value after gene filtering and log transformation and Max(expr)*=maximum expression value after gene filtering and log transformation.

2.1.1 Datasets used in DE gene detection and classification

In differentially expressed (DE) gene detection analysis and classification, datasets are analyzed to search for differentially expressed genes in patients with two types or multi-class and to identify molecular biomarkers of disease classification and prediction to diverse tumor types. (Golub, et al., 1999) with two leukemia (ALL and AML), (Alon, et al., 1999) of colon tumor, and two prostate cancer datasets of (Yu, et al., 2004)and (Luo, et al., 2001) are analyzed. And four survival datasets are assessed to identify differentially expressed (DE) genes linked to survival outcome. One is (Singh, et al., 2002) including 102 prostate tissue samples, 52 tumor samples with 8 recurrent and 13 non-recurrent with survival time. Another is (Beer, et al., 2002) with 86 primary lung adenocarcinomas, including 67 stage I and 19 stage III tumors, and 10 non-neoplastic normal lung samples. And Stage I and III were also grouped into high-risk and low-risk subgroup, respectively. Another is prostate cancer dataset with relapse survival information by (Yu, et al., 2004). It contains including 23 organ donors normal samples and 66 tumors. Another is van'tVeer dataset collected in Nature 2002 by (van't Veer, et al., 2002). The data set is composed of 97 primary breast cancers including 46 from patients who developed distant metastases within 5 years, 51 from patients who continued to be disease-free after a period of at least 5 years. For above four survival data sets with multi-class, we only select two sub groups from full samples in original data. We summarize the descriptions of original datasets and subgroups selected from original data sets used in DE gene detection and classification more details in Table 1 and following sections more details.

A. Golub (GOL) data

This Leukemia dataset (Golub, et al., 1999) is one of the most well known data set for methodological development. It contains 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of actue myeloid leukemia (AML) samples which are the combined training samples (38 samples: 27 ALL, 11 AML) as the primary samples and test samples (34 samples: 24 bone marrow and 10 peripheral blood samples) as the independent samples. The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7129 human genes are collected initially. And then we deleted negative and zero gene expression values and then take log transformation. Finally, 1994 genes and 72 samples remained. The dataset has 0 and 16.1230 gene expression value as the minimum and maximum, respectively.

B. Alon (ALO) data

This dataset is originally collected in PNAS (1999) by (Alon, et al., 1999). The data matrix of Affymetrix oligonucleotide array contains about 6500 features and 62 samples. Some of the features display a hybridization signal that is many times stronger than their neighbors (~ 4% of the intensities are > 3 SD away from the mean for expressed tags(ESTs) as mentioned in the original paper. These outliers are deleted. To compensate of each EST on an array was normalized by dividing with the mean intensity of all ESTs on that array and multiplying with a nominal average intensity (50). The expression values of 2000 genes and 62 samples of 40 tumor and 22 normal colon tissues are finally remained. The dataset was downloaded on the web at http://www.molbio.princeton.edu/colondata.

C. Luo (LUO) data

The data were originally published in Cancer Research by (Luo, et al., 2001). The data included 16 prostate adenocarcinoma samples from Johns Hopkins Hospitals during October 1998 and March 2000, and 9 benign prostatic hyperplasia specimens from Johns Hopkins Hospital during February 1999 and November 2000. Total 25 samples and 6500 human genes were analyzed using cDNA microarray. We deleted some genes with zero and negative gene expression values, and took log2 transformation. Finally, 6433 genes and 25 samples remained.

D. Singh (SIN) data

The data set was originally published in Cancer Cell by (Singh, et al., 2002). The initial data set is composed of 12,600 genes and 102 samples including 52 prostate tumors and 50 non-tumor prostate samples using oligonucleotide microarrays. In the original dataset, it is already log2-transformed dataset. However, there exist too many zero values. After filtering out gene with zero values, 1662 genes remained for the further analysis in this study. In the original paper, this dataset indicated that 317 genes had up-regulated in the tumor samples and 139 genes had up-regulated in the normal samples. Additionally, this dataset has 21 patients with respect to recurrence following surgery with 8 patients having relapsed and 13 patients having remained relapsed free for at least 4 years.

E. Lapointe (LAP) data

The prostate cancer dataset was originally published in PNAS by (Lapointe, et al., 2004). It contains 39009 genes and 112 samples with 62 prostate cancers, 41 normal samples, and 9 metastasis prostate cancer samples in lymph node. This original dataset has log2 transformation and some of missing values. After filtering out genes with missing values and additional filtering criterion, 3098 genes remained finally. Here the used specific criterion is to delete some more genes with low gene expression values such as mean < 1.5 and standard deviation < 1.2 for gene expression value in each gene.

F. van’tVeer (VAN) data

The data were collected in Nature 2002 by (van't Veer, et al., 2002). The data set is composed of 97 primary breast cancers including 46 from patients who developed distant metastases within 5 years, 51 from patients who continued to be disease-free after a period of at least 5 years. From the raw dataset, after eliminating gene with missing values and low expression values of less than log2 gene expression value 0.7,3176 genes are used for further analysis in this study.

G. Yu (YU) data

The data set was originally published in JCO 2004 by (Yu, et al., 2004). The initial data set contains 152 samples, which contain 89 samples with 66 primary prostate cancer samples and

23 organ donor normal samples and 37777 genes. All samples were analyzed using Affymetrix U95A microarray chips. In the original paper, they deleted some genes whose expression was very similar throughout all the samples to maximize the difference between the three groups (PC: prostate cancer, OD: donor prostate, AT: prostate tissues adjacent to cancer) and eliminated genes with low gene expression value less than the arbitrary cut-off value. And 19139 genes remained. After data preprocessing in which is filtered out genes with negative (or zero) gene expression value and missing values and log transformation are taken respectively. Finally, 2532 genes remained in the dataset for further analysis.

H. Beer (BEE) data

The data were originally published in Nature Medicine 2002 by (Beer, et al., 2002). The 86 lung adenocarcinoma samples were collected from the University of Michigan Hospital between May 1994 and July 2000 from 67 stage I and 19 stage III patients, and 10 non-neoplastic lung tissues were also obtained during that time. The total 96 samples were analyzed using Affymetrix HG6800 microarray chips. From the raw data set, after deleting negative and zero expression values and taking log transformation, 3577 genes remained in the dataset. In DE gene detection and classification, 86stage I and 10 non-neoplastic lung tissues are used.

2.1.2 Datasets used in Cluster analysis

In gene clustering, all data are pre-filtered (pre-processing) by a criterion to eliminate genes with negative and zero expression value.

A. Spellman-alpha (SP.AFA) data

This data set is originally collected to create a comprehensive catalog of yeast genes whose transcript levels varied periodically within the cell cycle by (Spellman, et al., 1998). DNA microarray samples from yeast cultures were synchronized by alpha factor arrest. It is a time series and cyclic data. After pre-processing, 4480 genes and 18 samples remained.

B. Spellman-elu (SP.ELU) data

This data set is also collected in(Spellman, et al., 1998). The only difference from Spellman-alpha data is the differently synchronized yeast by elutriation. After pre-processing, 5766 genes and 14 samples remained.

C. Causton (CAU) data

This data setis originally collected by (Causton, et al., 2001)to explore how gene expression in Saccharomyces cerevisiae is remodeled in response to various changes in extracellular environment, including changes in temperature, oxidation, nutrients, pH, and osmolarity. After pre-processing, 4616 genes and 45 samples remained.

2.2 MV METHODS DESCRIPTION

In microarray expression data, missing values frequently occur due to diverse sources, such as scratches, dust, insufficient resolution and hybridization error, etc on arrays. However, unfortunately, most of statistical analyses require a complete matrix as the input. Thus, estimating missing values are important as they affect downstream analyses such as differentially expressed gene detection analysis, k-means clustering, and classification. One simple strategy is to delete genes with missing values and keep a complete matrix. However, this may lead to a loss of large useful information. Especially, it is rarely to have a set of complete values over all experiments. Therefore, in microarray experiment, more commonly suggestible strategy is to estimate missing values by borrowing the information of genes with similarity structure. During the last decade, although various imputation methods have been developed and proposed, as any given dataset, uniformly superior imputation is still ambiguous because each imputation method has own strength and weakness and imputing missing value largely depends on nature and structure of dataset. Some of imputation methods have a good performance on local structure, while others show better performance on global data structure. Therefore, in this comparative study, we employ 10 imputation methods including global and local imputation method.

2.2.1 Naïve methods

The most naïve way to impute missing values in microarray data is to insert the corresponding row/column averages or zero value. Such methods are highly inaccurate. In addition, (Troyanskaya, et al., 2001) already showed that row/column average method yields poor performance comparing to K-nearest neighborhood imputation method described below. We will omit the naïve methods in all analyses hereafter and focus on the more robust imputation methods.

2.2.2 K-nearest neighbor (KNN) based on distance/correlation

K-nearest neighbor introduced in (Troyanskaya, et al., 2001)finds the most similar k genes for the target gene with missing values based on euclidean distance (KNN.e) or pearson correlation measure (KNN.c). The method was among the first proposed methods and became popular due to its simplicity. The following steps show the algorithm for imputing the missing values in a given gene g,

Step1: Compute the Euclidean distance (or Pearson correlation) between g-th gene and all remaining genes using only those co-ordinates not missing in g-th gene. Select the most similar K genes.

Step2: Impute the missing entries of g-th gene by weighted averaging the corresponding coordinates of the K genes. The weights are chosen to be the inverse of distances or correlations.

Previous literatures suggested that although the result depends on K, the range of 10 to 20 provides good and the consistent imputation results. We will use K=10 in this thesis.

2.2.3 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) introduced by (Troyanskaya, et al., 2001) approach first imputes all missing values using the row average imputation method in a preliminary step since SVD cannot handle the data matrix with missing values. It is then applied to create a set of mutually orthogonal principal components of expression patterns, so-called eigen-genes. SVD is then linear transformation of the expression data (A) from G genes × S samples (arrays) to the reduced $p \times S$ eigen-arrays, where p is the proportion of eigen-genes which correspond to the largest eigen-values are selected values to reconstruct MVs in the expression matrix.

The singular value decomposition of A is

$$A = U\sum V^T \tag{7}$$

The columns of V^T form the eigen-genes of $A^T A$, whose contribution to the expression in the eigen-space is quantified by corresponding eigen-values on the diagonal of matrix $\sum$. The k most significant eigen-genes are selected to form the basis for the imputation process. The value of k is usually determined empirically. The missing entry is estimated from a linear combination of the k eigen-genes weighted by the regression coefficients. This process is iterated until the total change in the matrix A converges to a sufficiently small prefixed value. In previous literature, a range of 0.1~0.25 for p gives good and consistent results. We will use p=0.15 in this study.

2.2.4 OLS

Ordinary least squares imputation method introduced by (Nguyen et al., 2004) is based on local neighboring-based approach as KNN. While KNN is to impute a missing value by using a weighted average of K most similar genes, OLS is regressed over K most similar genes. Therefore, a missing value is imputed by the weighted average of predicted values of fitted regression of the gene with missing values onto each neighbor gene, where K most similar genes are selected by absolute Pearson correlation value and the weight is,

$w = \left(\frac{r_{\gamma x}^2}{1 - r_{\gamma x}^2 + 10^{-6}} \right)^2$, where $r_{\gamma x}$ is the correlation between the target gene (γ) with MVs and the candidate gene (x). We will use K=10 in this study.

2.2.5 PLS

Repeated ordinary least squares a neighboring-based approach proposed by (Nguyen et al., 2004) and Bo et al. (2004) are to regress missing value over each of the k most similar neighbor genes as mentioned 2.3.3 part. Succinctly, missing values (MVs) are estimated through the weighted average of the predicted values from the regression of the target gene with missing values onto each neighbor gene. K-closest genes are selected by neighbor genes with largest absolute Pearson correlation value from the candidate gene set C_s . Nguyen et al introduced a novel method Partial Least Squares (PLS). In this paper, they concluded that PLS has the strength to have uniformly better performance in view of accuracy across the wide range of observed gene expression, whereas KNN imputation method presents worse performance over some specific ranges of expression values.

The main difference of PLS (partial least squares) by Nguyen et al. (2004) from KNN is to use all the candidate gene expressions as well as the available values from the target gene to estimate missing values (MVs). More precisely, the procedure is like below,

Step1: PLS constructs a sequence of gene components with the candidate gene expression matrix and the available values of the target gene.

Step2: The target gene and candidate genes comprise missing values (MVs) and available values, respectively. Denote the expression values of the target gene as $y_g = \begin{pmatrix} y_g^A \\ y_g^M \end{pmatrix}$, where y_g^A and y_g^M are available values and missing entries to be imputed of target gene g. And the expression values of candidate genes for the target gene g are denoted by $Y_{-g} = \begin{pmatrix} Y_{-g}^A \\ Y_{-g}^* \end{pmatrix}$, where Y_{-g}^A is a $G_g \times S_g$matrix of available values corresponding to y_g^A and Y_{-g}^* comprises of available values corresponding to the MVs of target gene g, y_g^M . Thus, PLS contains a training set (y_g^A , Y_{-g}^A) and the test set Y_{-g}^* will be used to predict MVs (y_g^M) of target gene g.

Step3: Since the number of samples (S_g) is much smaller than the number of available genes (G_g) ($S_g << G_g$), dimension reduction is necessary. Thus, PLS imputation is a dimension reduction method,which extracts gene components sequentially to maximize the sample covariance between the target gene and the linear combination of the set of candidate genes. The more detailed procedure of dimension reduction is explained in the following steps,

Step4: The k-th PLS step seeks a weight vector, $w_k(g)$ $(\mathrm{G_g} \times 1)$, such that $w_k(g) = \underset{w'w=1}{\arg\max} \operatorname{cov}^2(Y_{-g}^A w, y_g^A)$, subject to the orthogonality constraints $w_k^{'}(g) S w_d(g) = 0$, $\forall_d$, $1 \le \mathrm{d} \le \mathrm{k}$ where $S = Y_{-g}^{A'} Y_{-g}^A$. Thus, a sequence of weight $w_1(g)$, ...,$\mathrm{w}_2(g)$ are obtained from this step for each gene g with missing values (MVs). And according to the previous literature by (Brock, et al., 2008), the number of components is confident between 2 to 25. We will use K=10 in this thesis.

Step5: The PLS gene components of linear combinations with maximum covariance with target gene g, $t_k^A(g) = Y_{-g}^A w_k(g)$ are computed. Therefore, PLS imputation captures the most important mode of covariance exhibited between the target gene and candidate genes first and the next most important mode is captured to be orthogonal to the first PLS components.

Step6: Using the constructed PLS gene components as predictors, a linear regression model based on the available values is fitted. $\hat{y}_g{}^A = T^A(g)\hat{\beta}_g$, where $T^A(g)$ is a matrix of the K_G PLS gene components and $\hat{\beta}_g$ is the least squares regression coefficient estimates.

Step7: Next the test data is applied. That is, the expression values of candidate genes (Y_{-g}^*), corresponding to missing entries of target gene (y_g^M) are used to construct the test PLS components, $T^*(g)$ based on only the training information (Step4 and Step5). That is, the test components are substituted into the training PLS regression model to predict the MV, $\hat{y}_g^M = T^*(g)\hat{\beta}_g$.

2.2.6 LSA-impute

LSA-impute introduced by (Bo, et al., 2004) is to estimate missing values based on least squares principle as utilizing correlations between both genes and arrays. They are called LSimpute_gene and LSimpute_array, respectively. The gene-based estimates are obtained by multiple OLS method using the K closest candidate genes, and the array-based estimates are attained by multiple-regression based on the arrays, where missing values in gene expression matrix are substituted by gene-based estimates initially. The K closest genes are selected by the absolute Pearson correlation values. Thus, LSA imputation is the combined method of gene-based and array-based imputation estimates. In the paper, there are two variants of estimate combination. The first (LSimpute_combined) is to use a fixed global weighting of the estimates between LSimpute_gene and LSimpute-array. The best global weight of two estimates is determined by initially re-estimating from the known values in the gene expression matrix and minimizing the sum of errors for re-estimated data.

2.2.7 LLS-impute

LLS-impute proposed by (Kim, et al., 2006) is based on local least squares principle to represent a target gene with missing values as a linear combination of K coherent genes that have the large absolute Pearson correlation values from candidate gene set. As in OLS and LSA-impute, the LLS-impute is to estimate missing values by performing multiple regressing the candidate genes on target gene. However, the least squares estimates are determined by pseudo-inverse of K closest genes, where if the K closest genes have some missing values and its percentage is relatively small, then K neighboring genes are deleted in determining estimates, otherwise MVs are initially estimated by row-average imputation method.

2.2.8 BPCA

This approach is based on Bayesian principal component analysis (BPCA) proposed by (Oba, et al., 2003) to consist of three elements, (1) Principal component regression, (2) Bayesian estimation, (3) an expectation-maximization (EM)-like repetitive algorithm. It proceeds iteratively by switching between updating the posterior distribution of the PCA parameters and the posterior distribution of the missing values. The posterior distribution of model parameters is iteratively estimated until convergence is reached. This approach is a time consuming method because no parameters have to be fixed as the algorithm itself sets up the appropriate PCA dimension. Moreover, since BPCA considers the global correlation structure of the data, this algorithm may not be well suited for data, which has a dominant local correlation structure.

2.3 DOWN-STREAM ANALYSES EVALUATED

To evaluate biological impacts of missing value imputation in down-stream analyses, we consider three types of analyses commonly seen in microarray; differentially expressed (DE) gene detection, classification and gene clustering. The specific methods evaluated are described below.

2.3.1 DE gene detection

A two-sample microarray experimental design aims at identifying differentially expressed (DE) genes between two different groups such as normal versus disease. Various statistical tests have been employed to identify DE genes. The two t-test statistics provides a simple statistical method for identifying DE genes. But t-test statistics is unstable when the sample size is small, which causes an increase in false discovery rate. Moreover, the other weakness has the problem of multiple testing. Thereby, we employ adjusted Benjamini-Hochberg (BH) t-test statistics and LIMMA method to cope with the problem of multiple testing and SAM method is used to handle small variance by adding a small positive constant to the denominator. Thus, we will evaluate the impact of imputation method on down-stream analyses using three popular methods, Benjamini-Hochberg adjusted p-value from t-test, SAM, and LIMMA to identify differentially expressed genes. FDR is controlled at 5% and the default parameters are used in the packages.

2.3.1.1 T-test + Benjamini-Hochberg (BH) adjusted p-value

Standard statistical method, t-statistics is applied here to compare treatment group versus normal group, or other conditions with two classes.

$$t_g = \frac{\bar{x}_1 - \bar{x}_2}{s_{x_1x_2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\text{, where } S_{x_1x_2} = \sqrt{\frac{\sum(x_i-\bar{x}_1)^2+\sum(x_j-\bar{x}_2)^2}{n_1+n_2-2}} \qquad (8)$$

$\bar{x}_1$ and $\bar{x}_2$ are defined as the average levels of expression for g-th gene in class 1 and 2. $s_{x_1x_2}$ represents the standard deviation of repeated expression measurements and n_1 and n_2 are the numbers of measurements in class 1 and 2. Since microarray experiments simultaneously monitor expression levels of thousands of genes, there is multiple comparison issue. To resolve the issue, many approaches have been introduced for adjusting multiple testing so far. Especially (Benjamini and Hochberg, 1995) adjusted p value is a popular step-up procedure which controls the FDR (false discovery rate) defined as $E\left[\frac{\text{false rejections}}{\text{total rejections}}\right]$, where if there are no rejections in study, then it is defined by 0. Thus, if we control FDR at 0.05, then we can claim on average, no more than 5 % of the rejections are in error under some dependency structures with $p_{r_j}{}^{*} = \min_{k=j...m}\{\min(\frac{m}{k}p_{r_k},1)\}$. More precisely, the BH procedure is like below, Consider testing $H_1, H_2, \ldots, H_m$ based on the corresponding p-values, $P_1, P_2, \ldots, P_m$, where each hypothesis is that there is no difference between class 1 and class2 for each gene.

Step1: Let $P_{(1)} \le P_{(2)} \le \ldots \le P_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$.

Step2: Step up. Compare $p_{(k)}$ to $\frac{k \times \alpha}{k}$. If $p_{(k)} \le \frac{k \times \alpha}{k}$, then reject all of $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ and stop.

Step3: Compare $p_{(k-1)}$ to $\frac{(k-1) \times \alpha}{k}$. If $p_{(k-1)} \le \frac{(k-1) \times \alpha}{k}$, then reject all of $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$ and stop.

Step4: Compare $p_{(k-2)}$ to $\frac{(k-2) \times \alpha}{k}$. If $p_{(k-2)} \le \frac{(k-2) \times \alpha}{k}$, then reject all of $H_{(1)}, H_{(2)}, \ldots, H_{(k-2)}$ and stop.

Otherwise continue. Continue in this fashion until a stop or until no hypotheses are rejected.

Hence, adjusted p-values for Benjamini-Hochberg are shown below. It is a step-up method so we start from the opposite side.

$$p_{r_k} = \frac{\mathrm{k} \times \mathrm{p}_{(\mathrm{k})}}{k}$$

$$p_{r_{k-1}} = \min(p_{r_k} , \frac{\mathrm{k} \times \mathrm{p}_{(\mathrm{k-1})}}{k-1})$$

$$p_{r_{k-2}} = \min(p_{r_{k-1}} , \frac{\mathrm{k} \times \mathrm{p}_{(\mathrm{k-2})}}{k-2})$$

….

$p_{r_2} = \min(p_{r_3} , \frac{\mathrm{k} \times \mathrm{p}_{(2)}}{2})$

$p_{r_1} = \min(p_{r_2} , \frac{\mathrm{k} \times \mathrm{p}_{(1)}}{1})$

And then we compare the adjusted p-values to α. All of the adjusted p-values that are less than α correspond to rejections of null hypotheses using the given method.

T-test adjusted by BH procedure has the merit of fast computation and can be performed in excel without the need of programming. It, however, is usually less powerful and the t-statistic may be problematic when the variance component in the denominator is close to zero. SAM has been proposed to cope with unstable variance by adding a positive constant in denominator.

2.3.1.2 SAM (Significance Analysis of Microarray)

Significance Analysis of Microarrays (SAM) by (Tusher, et al., 2001)computes a score to each gene on the basis of change in expression relative to the standard deviation of repeated measurements. SAM method adds a small positive constant to denominator of t-statistics as a fudge factor to avoid identifying falsely significant genes due to small variances. That is, if g-th gene has low expression values, variance in $t_{sam}(g)$can be high, due to small values of$s_{(g)}$. However, to compare $t_{sam}(g)$across all genes, the distribution of $t_{sam}(g)$should be independent of the level of gene expression and of $s_{(g)}$. Thus we choose a fudge factor, s_0 to make the coefficient of variation of $t_{sam}(g)$approximately constant as a function of $s_{(g)}$ by the similar approach as Efron et al. In other words, s_0 constant is chosen to minimize the coefficient of variation.

$$t_{sam}(g) = \frac{\overline{x}_1 - \overline{x}_2}{s_{x_1 x_2}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} + s_0},$$

where $s(g) = S_{x_1 x_2}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} = \sqrt{\dfrac{\sum(x_i - \overline{x}_1)^2 + \sum(x_j - \overline{x}_2)^2}{n_1 + n_2 - 2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}$ (9)

SAM provides FDR (false discovery rate) to be calculated by Median (or 90th percentile) of # of falsely called genes by dividing # of genes called significant.

More precisely, the SAM algorithm is stated as

Step1: Order test statistics in equation 9 according to magnitude.

Step2: Based on null hypothesis that there is no difference between class 1 and class 2, Using permutation test, for each permutation, compute the ordered null (unaffected) scores.

Step3: Plot the ordered test statistic against the expected null scores.

Step4: Call each gene significant if the absolute value of the test statistic for that gene minus the mean test statistic for that gene is greater than a stated threshold.

Step5: Estimate the false discovery rate based on expected versus observed values.

2.3.1.3 LIMMA

It is a tool for identifying differentially expressed genes involving comparisons between two groups proposed by (Smyth, 2004). The main idea is to fit a linear model to the expression data for each gene. Empirical Bayes and other shrinkage methods are applied to borrow prior information across genes making the analyses stable even for experiments with small sample size. In the model, there are three main steps. The first step is to rearrange it in the structure of general linear models with arbitrary initial coefficients and contrasts of interest. The second step is to derive consistent and robust closed form estimators for hyper-parameters even for small sample size based on the marginal distributions of the observed statistics. The third step is to reformulate the posterior odds statistics in terms of moderated t-statistics in which posterior residual standard deviations are used in place of ordinary standard deviations. In the end, those steps make it possible to have more stable inference when even sample size is small as the approach proposed by Lonnstedt and Speed. This package then provides the B-H adjusted p value after multiple testing.

2.3.2 Classification

We performed classification with feature selection based on univariate t statistics using LDA, KNN (k=5), and SVM classifier with linear kernel function. Univariate method considers one variable (a feature) at a time, whereas multivariate method considers subsets of variables (features) together such as PAM. We adopted the simple univariate feature selection by t statistics. Weperformed leave one out cross validation (LOOCV), selected the top N=5, 10, 30,50, 100 gene features with the largest t-statistics and picked the one that generates the largest

Youden Index(smallest error rate). For PAM, gene selection is embedded and we pick the threshold that generates the best accuracy.

2.3.2.1 LDA classifier

In Linear Discriminant Analysis (LDA) proposed by (Fisher, 2000), each class is characterized by its vector of means or 'centroid'. An unknown sample is evaluated by computing the scaled distance between its expression profile and each class centroid. The unknown is assigned to the class to which it is nearest. Thus, LDA can be thought of as a nearest centroid classifier. The procedure of LDA is described more precisely below.

We would like to classify unknown samples into one of K classes. To build a classifier, we obtain n_k training samples per class, k=1,2,…,K, with g genes on each microarray. For each training sample, we observe class membership, sample information X and expression profile Y. For simplicity, we will utilize only two classes (1 or 2) in this study. Note that each expression profile is a vector of length m. We assume that expression profiles from class K are distributed as N (μ_k,Σ), the multivariate normal distribution with mean vector μ_k and covariance matrix Σ. Call L ($\cdot$; μ_k,Σ) the corresponding probability density function. Finally, we agree upon prior probabilities π_k that an unknown sample comes from class k, k=1 and 2. Bayes' theorem states that the probability that a sample comes from class k, given that sample's expression profile, is proportional to the product of the class density and prior probability:

$$\Pr(Y = k \mid X = x) \propto L(x;\mu_k,\Sigma) \times \pi_k \qquad (11)$$

We call equation (11) the posterior probability that array x comes from sample k. LDA assigns the sample to the class with the largest posterior probability:

$$\hat{y} = \arg\max_k \left\{ L(x;\mu_k,\Sigma) \times \pi_k \right\} \qquad (12)$$

This can be shown to be the rule that minimizes misclassification error by Mardia et al. (1979).

The innards of the right side of equation (12) are proportional to

$$|\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \log(\pi_k) \right\} \qquad (13)$$

Since the covariance matrix ∑ is the same for all classes, only the exponential component of equation (14) is relevant to classification. We can then rewrite equation (12) as

$$\hat{y} = \arg\max_k \left\{ (x-\mu_k)^T \Sigma^{-1}(x-\mu_k) - 2\log(\pi_k) \right\} \qquad (14)$$

Thus, a sample is assigned to the class to which it is nearest, as measured by the metric $\|x-\mu\|^2 - 2\log(\pi)$ is the square of the Mahalanobis distance between x and μ. We can further simplify the problem by assuming independence between genes. This allows us to simplify the LDA classification rule (14) to

$$\hat{y} = \arg\max_k \left\{ \sum_{i=1}^{m} \left(\frac{x_i^* - \mu_{ik}}{\sigma_i} \right)^2 - 2\log(\pi_k) \right\} \qquad (15)$$

2.3.2.2 KNN classifier

KNN classifier in (Belur V. Dasarathy., 1991) is based on a distance/similarity function for pairs of observations, such as the Euclidean distance. K nearest neighbors of a training data is computed first. We fixed K=5 KNN classification by selecting from an exploratory examinations of optimal parameter selection to perform some tests using 3 CD (complete data set), GOL, ALO, and LUO with k=1 to k=15. Then the similarities of one sample from testing data to the k nearest neighbors are aggregated according to the class of the neighbors, and the testing sample is assigned to the most similar class. A major drawback of the similarity measure used in KNN is that it uses all features equally in computing similarities. It can lead to poor similarity measures and classification errors, when only a small subset of the features is useful for classification. Therefore, in KNN, confident feature selection is suggested.

2.3.2.3 SVM classifier

SVM introduced in (Burges, 1998)provides a machine learning algorithm for classification Gene expression vectors can be thought of as points in an n-dimensional space. The SVM is then trained to discriminate between the data points for that pattern (positive points in the feature space) and other data points that do not show that pattern (negative points in the feature space). Specifically, SVM chooses the hyper-plane that provides maximum margin between the plane surface and the positive and negative points. The separating hyper-plane is optimal in the sense that it maximizes the distance from the closest data points, which are the support vectors. The mathematical background of SVM is like below, given a training set of instance-label pairs $\{x_i, y_i\}, i = 1,\dots,l$, where $x_i \in R^n$ and $y \in \{1,-1\}^l$, thus, we consider the problem of separating the set of training vectors belonging to two separate classes. The support vector machines (SVM) require the solution of the following optimization problem:

$$\min_{w,b,\varepsilon} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \varepsilon_i \text{, subject to } y_i (w^T \varphi(x_i) + b) \geq 1 - \varepsilon_i \text{, } \varepsilon_i \geq 0 \qquad (20)$$

Training vectors x_i are projected into a higher dimensional space by the function φ. Here there are many possible linear classifiers that can separate the given data, but the only one to maximize the margin exists. In other words, SVM finds a linear separating hyper-plane to maximize the distance between it and the nearest data point of each class. $C > 0$ is the penalty parameter of the error term. And then the various kernel functions such as linear, polynomial, radial basis function, sigmoid, and etc are applied with the equation (21). Each kernel function with the kernel parameters $\gamma, r,$ and d is explained in the following.

$$K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \tag{21}$$

1) Linear: $K(x_i,x_j) = x_i^T x_j$
2) Polynomial: $K(x_i,x_j) = (\gamma x_i^T x_j + r)^d$, $\gamma > 0$.
3) Radial basis function (RBF): $K(x_i,x_j) = \exp(-\gamma\|x_i - x_j\|^2), \gamma > 0$.
4) Sigmoid: $K(x_i,x_j) = \tanh(\gamma x_i^T x_j + r)$.

2.3.2.4 PAM

The (Tibshirani, et al., 2002) uses the statistics $d_{gk} = \dfrac{\overline{x}_{gk} - \overline{x}_g}{w_k(s_g + s_0)}$ to select genes, where $w_k = (1/n_k - 1/n)^2$ makes $w_k \times s_g$ equal to the standard error of the numerator, and s_0 is a fudge factor intended to guard against very large statistics for very small standard errors as like SAM method; by default, PAM chooses the median of the s_g for s_0. Without s_0, d_{gk} is just a t-statistics comparing the mean of gene g in class k (1 or -1, e.g normal or disease) with the overall mean of gene g. Hence, d_{gk} measures the difference between gene g in class k and gene g in all classes combined. A gene that discriminates one class from the rest will have a statistic of large absolute value. PAM then shrinks d_{gk} toward zero, eliminating the genes that do not provide sufficient discriminatory information. For a particular choice of shrinkage parameter Δ, the shrunken statistics is

$$\tilde{d}_{gk} = sign(d_{gk})(\left|d_{gk}\right| - \Delta)_{+}, \text{ where'+' means 'positive part'} \qquad (22)$$

Thus, all d_{gk} less than Δ in absolute value are shrunken to zero, and the rest are shrunken to somewhere between zero and their original values. The shrinkage of the remaining statistics toward zero is intended as a 'de-noising' step. We can then reformulate equation (23) with the shrunken statistics to produce corresponding shrunken centroids,

$$\tilde{x}_{gk} = \tilde{d}_{gk} \times w_k \times (s_g + s_0) + \bar{x}_g,$$ where shrinkage is of the class centroids toward the overall centroid. (23)

The genes for which all shrunken class statistics $\tilde{d}_{g1},\ldots,\tilde{d}_{gK}$ equal zero have shrunken centroid components that equal the corresponding components of the overall centroid. When distances from a new sample to the shrunken class centroids are computed in equation (15), the components for these inactivated genes are identical for each class. Distances are assessed as in equation (15), with the shrunken centroid $\tilde{x}_{gk}$ replacing μ_{gk} and $s_g + s_0$ replacing σ_g . Any prior information on class prior weights on each class in proportion to its sample prevalence can be induces in π_k . One simple choice is $\pi_k = n_k / n$, placing prior weights on each class in proportion to its sample prevalence; another is π_k =1/K, placing equal prior weights on each class. Briefly, a PAM classifier selects $\tilde{m}$ genes by soft thresholding based on the modified t-statistics, d_{gk} , then shrunken statistics to update the class centroids. The classifier can be represented by the shrunken centroid components and pooled standard deviations of the active genes, since these are the only components needed in the distance function. Each centroid is now of length $\tilde{m}$, with g-th component somewhere between that gene's class and overall means. Nothing is known about the distribution of active genes across classes. The selected genes are interpreted as simultaneously distinguishing all classes from each other.

2.3.3 Cluster analysis

The objective of cluster analysis in microarray expression data is to group genes or experiments into clusters with similar patterns. A cluster analysis is called unsupervised learning in the view of the classes are unknown a priori and need to be discovered from the data unlike supervised learning such as classification and discriminant analysis. In this study, we included KNN and SOM. Since the number of clusters K usually cannot be determined for a given data set, we run gene clustering using different choices of K. Due to the already demanding computation, we only tested KNN and SOM with K=5, 10, and 15.

2.3.3.1 K-means

K-means clustering by (Stuart Lloyd., 1957) is one of the common unsupervised algorithms to define k-centroids, one for each cluster. This is a method of clustering that produces a partition of the data into a particular number of k groups set. From an initial partition, individuals are moved into other groups if they are closer to its mean vector that that of their current group (Euclidean distance measure is generally used here). After each move, the relevant cluster mean vectors are updated. The procedure continues until all individuals in a cluster are closer to their own cluster mean vector than to that of any other cluster. It seeks to minimize the variability within clusters and maximize variability between clusters. Finding the optimal number of groups will also be an issue in K-means clustering. As a preliminary test, we evaluated K values using gap statistics for Spellman-alpha, Spellman-elu, and Causton dataset. For all 3 datasets, gap-statistics does not present the difference too much for each cluster from 1 to 20 clusters. Therefore, we will perform k-means for K=5, 10, and 15. Thereby, we will not discuss the optimal K value in details here.

2.3.3.2 SOM

As a machine-learning method, a self-organizing map (SOM) proposed and studied by (Kohonen, 2001) belongs to the category of neural networks. It provides a technique to visualize the high-dimensional input data (in our case, the gene expression data) on an output map of neurons (also called nodes). The map is frequently in a two-dimensional grid (usually of hexagonal or rectangular geometry) of neurons. In the high-dimensional input space, the structure of the data is represented by prototype vectors serving similar functions as the mean vectors in the k-means algorithm, each of which is related to a neuron in the output space. As an input for the algorithm, the dimension of the output map (e.g., a map of 6 x 5 neurons) needs to be specified. After initializing the prototype vectors, the algorithm iteratively performs the following steps.

Step1: Every input vector representing a gene expression profile is associated with the closest prototype vector, and thus is also associated with the corresponding neuron on the output space.

Step2: Update the coordinates of a prototype vector based on a weighted sum of all the input vectors that are assigned to it. The weight is given by the neighborhood function (a kernel function in nature), which can be a Gaussian distribution function, applied in the output space. That is, in the updating step, a prototype vector is pulled more toward input vectors that are closer to the prototype vector itself and is less influenced by the input vectors located farther away. In the meantime, this adaption procedure of the prototype vectors is reflected on the output nodes- nodes associated with similar prototype vectors are pulled closer together on the output map.

Step3: To put a simulated annealing kind of flavor, the initial variance of the Gaussian neighborhood function is chosen so that the neighborhood covers all the neurons, but then the variance is decreased during an iteration so as to achieve a smoother mapping. The algorithm terminates when convergence of the prototype vectors is achieved. From the cluster analysis point of view, SOM methods look similar to K-means clustering method. SOM clustering differs from K-means clustering in that a cluster has “two faces” in an SOM. It is represented by the prototype vector in the input space and the neuron on the output space. In this way, an SOM provides a direct means to visualize relations among different clusters. Moreover, a prototype vector is adjusted according to not only the data points that are associated with it but also data points that are assigned to other prototype vectors.

2.4 QUANTITATIVE EVALUATION: ROOT MEAN SQUARED ERRORS (RMSE)

Variants of root mean squared errors (RMSE) are commonly used as a statistical quality to measure of how close the estimated values are to real values. Based on such RMSEs, the evaluation procedure of missing imputation methods are like following steps.

Step1: Given a real gene expression matrix, the (real) MVs are removed to form a complete gene expression matrix without MVs. We call this gene expression matrix CD (complete data). It is denoted by CD=$(y_{gs})_{G\times S}$.

Step2: Next, a proportion q, 0<q<1, of MVs are intentionally introduced by randomly removing values in CD, where q=0.01, 0.05, 0.1 and 0.2. Let us denote this gene expression matrix with MD (missing data).

Step3: Imputation methods are applied to estimate MVs in MD. For missing entries, missing value, $(y_{gs})_{G\times S}$in MD is substituted by imputed value ($\hat{y}_{gs}$). We call this gene expression matrix ID (imputed data).

Step4: Compare imputed values of ID to true values of CD to access the accuracy (performance) of each imputation method with respect to RMSE.

Step5: Repeat from step 2 to step 4 for each missing percentage and imputation method.

Step6: Take the average value for RMSE in step 5. It represents the performance of an imputation method when assuming a specific missing percentage.

Below 6 diverse RMSEs have been introduced in the papers of studies to missing imputation methods. Thus, Table 2 shows four variants of RMSE, RAE, and LRMSE used in the previous literature to evaluate MV imputation performance.

Table 2.Variant of RMSE

Literature	Used measure	Values used in equation
(Bo, et al., 2004)	RMSE	Original expression values
(Troyanskaya, et al., 2001)	NRMSE1	Original expression values
(Kim, et al., 2006; Oba, et al., 2003)	NRMSE2	Original expression values
(Ouyang, et al., 2004)	NRMSE3	Original expression values
(Nguyen et al., 2004)	RAE	Original expression values
(Brock, et al., 2008)	LRMSE	Log transformed values

(Bo, et al., 2004) used non-normalized score, RMSE (root mean squared error) between the true values and the estimated values in equation (1)

$$\text{RMSE} = \sqrt{\frac{1}{\#\ \text{of missingness}} \sum_{\{y_{gs}\ mis\sin g\}} (\hat{y}_{gs} - y_{gs})^2} \qquad (1)$$

(Troyanskaya, et al., 2001) normalized the RMSE by dividing it by a normalizing constant, the average value over all observations in the true full dataset, to compare the performance of imputation methods using different datasets with equation (2).

$$\text{NRMSE1} = \frac{\sqrt{\frac{1}{\#\ \text{of missingness}} \sum_{\{y_{gs}\ \text{missing}\}} (\hat{y}_{gs} - y_{gs})^2}}{\frac{1}{G \times S} \sum_{g} \sum_{s} y_{gs}} \qquad (2)$$

(Kim, et al., 2006; Oba, et al., 2003) normalized the RMSE by dividing it by a different normalizing constant, the standard deviation of the values in the true full dataset over missing entries with equation (3)

$$\text{NRMSE2} = \sqrt{\frac{\frac{1}{\#\,\text{of missingness}} \sum_{\{y_{gs}\text{missing}\}} (\hat{y}_{gs} - y_{gs})^2}{\frac{1}{(\#\,\text{of missingness} - 1)} \sum_{\{y_{gs}\text{missing}\}} (y_{gs} - y_{..})^2}} \qquad (3)$$

(Ouyang, et al., 2004) normalize the RMSE by dividing it by another different normalizing constant, the root mean square of original values of the missing entries with equation (4)

$$\text{NRMSE3} = \sqrt{\frac{\frac{1}{\#\ \text{of missingness}}\sum_{\{y_{gs}\,\text{missing}\}}(\hat{y}_{gs} - y_{gs})^2}{\frac{1}{\#\ \text{of missingness}}\sum_{\{y_{gs}\,\text{missing}\}}(y_{gs})^2}} \qquad (4)$$

The normalization of RMSE is a step to make it possible to carry out the comparison basically to have the level of difficulties of comparison to imputation method in variant datasets with different scale. For selecting the best MV imputation method in a given data sets, however, all the four RMSE variants converged an identical method by representing the same ranks for the performances of imputation methods in the preliminary results. Hereby, we will keep NRMSE by (Troyanskaya, et al., 2001) as a representative 3 NRMSEs, LRMSE by (Brock, et al., 2008), and RAE by (Nguyen et al., 2004) for further analyses.

(Nguyen et al., 2004) used RAE measure to compare various imputation methods. Unlike NRMSEs, it has a slight modification in the equation to eliminate some minor drawbacks when y_{ij} equals zero and small values in equation (5)

$$\text{RAE} = \frac{1}{\#\ \text{of missingness}}\sum_{\{y_{gs}\,\text{missing}\}}\frac{\left|\hat{y}_{gs} - y_{gs}\right|}{\Phi(y_{gs})}, \Phi(y_{gs}) = \begin{cases} |y_{gs}| & if\, |y_{gs}| > \varepsilon \\ \varepsilon & if\, |y_{gs}| < \varepsilon \end{cases} \qquad (5)$$

Intuitively RAE is a better measure as it penalizes less for genes with high expression level. For example, an MV imputation error of 100 for genes with expression level at 200 is huge, while the error of 100 becomes ignorable for genes with expression level of 2000.

More recently, (Brock, et al., 2008)suggested LRMSE in equation (6) when the expression intensities are all positive. It can be easily shown that LRMSE approximately equals to,

$$\text{LRMSE}=\sqrt{\frac{1}{\#\ \text{of missingness}}\sum_{\{x_{gs}\ \text{missing}\}}\left(\hat{x}_{gs}-x_{gs}\right)^{2}}\ \text{,where}\ x_{gs}=\log(y_{gs})\ ,\ \hat{x}_{gs}=\log(\hat{y}_{gs}) \quad (6)$$

It is easy to show that LRMSE is an approximation of a square root of L_2-norm version of RAE:

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

In this thesis, we will evaluate the accuracy of 8 imputation methods using RMSE, LRMSE, and RAE.

2.5 BIOLOGICAL IMPACT EVALUATION: QUANTITATIVE MEASURES TO REFLECT BIOLOGICAL IMPACTS

Even though RMSE is a good measure in that it evaluates the differences of imputed values to the original values, the quantity brings a concern of not considering the ultimate biological impacts of missing value imputation to down-stream analyses. Recently, (Jornsten, et al., 2005) examined how missing values and their imputation affect significance analysis of differentially expressed (DE) genes using

$$\text{FPR}=\frac{\textit{false}\quad \textit{positive}}{\textit{true}\quad \textit{positive}\ +\ \textit{false}\quad \textit{positive}}$$

Similarly, (Scheel, et al., 2005) studied the influence of imputation on the detection of differentially expressed genes from cNDA microarray data using the percentage lost differentially expressed genes posterior MANOVA and SAM. Furthermore, (Tuikkala, et al., 2006) showed even when there are marked differences in terms of NRMSE across the datasets, these differences become negligible when the methods are evaluated in terms of how well they can reproduce the original gene clusters. They focused on assessing the agreement with the original clustering results when they performed clustering with estimated values from imputation methods. In this study, we will investigate the impact of missing value imputation on DE gene detection, classification and gene clustering analysis. Below we propose three biological impact measure for each of the down-stream analysis: biomarker list concordance index (BLCI) for DE gene detection, Youden's Index by (Youden, 1950) for classification and adjusted Rand index (ARI) by (Hubert ., 1985)for gene clustering,.

2.5.1 Biomarker list concordance index (BLCI) for DE gene detection

Suppose a complete data set (CD) is given. For evaluation purpose, missing values are randomly generated and imputed by a given MV imputation method. The imputed data is denoted as ID. 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 generated from ID (denoted as G_{ID}). We define the biomarker list concordance index (BLCI) between G_{CD} and G_{ID} as

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

, where *n(•)* is the number of genes of a given gene set, G_{CD}^{C} is the complement 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 the same as specificity. BLCI is equivalent to the famous Youden's index, which is sensitivity+specificity-1 by definition. We should note that taking the biomarker list from complete data as gold standard is necessary since we do not know the true biomarker list of the data. A higher BLCI value indicates that the biomarker lists from complete data and imputed data are similar and missing value imputation brings in smaller impact in down-stream biomarker detection. The simulations to generate BLCI evaluations are as follows.

Step1: Given a real gene expression matrix, the (real) MVs are removed to form a complete gene expression matrix without MVs. We call this gene expression matrix CD (complete data). It is denoted by CD=$(y_{gs})_{G\times S}$.

Step2: Next, a proportion q, 0<q<1, of MVs are intentionally introduced by randomly removing values in CD, where q=0.01, 0.05, 0.1 and 0.2. Let us denote this gene expression matrix with MD (missing data).

Step3: Imputation methods are applied to estimate MVs in MD. For missing entries, missing value, $(y_{gs})_{G\times S}$in MD is substituted by imputed value ($\hat{y}_{gs}$). We call this gene expression matrix ID (imputed data).

Step4: Perform DE gene detection analysis on CD and ID to generate their corresponding biomarker lists: G_{CD} and G_{ID}. Calculate *BLCI*(G_{CD}, G_{ID}).

Step 5: Repeat step2-4 for *N* times. We use *N*=100 in this thesis.

2.5.2 Youden's Index (YI) for Classification

Similarly we utilize Youden's index introduced by (Youden, 1950)as a quantitative measure to identify the impact of missing values in classification. Since we know the true class labels of the samples in this supervised learning scenario, we can directly evaluate the Youden's index of the prediction result from each imputed data. Specifically, we perform

Step1: Given a real gene expression matrix, the (real) MVs are removed to form a complete gene expression matrix without MVs. We call this gene expression matrix CD (complete data). It is denoted by CD= $(y_{gs})_{G\times S}$

Step2: Next, a proportion q, 0<q<1, of MVs are intentionally introduced by randomly removing values in CD, where q=0.01, 0.05, 0.1 and 0.2. Let us denote this gene expression matrix with MD (missing data).

Step3: Imputation methods are applied to estimate MVs in MD. For missing entries, missing value, $(y_{gs})_{G\times S}$in MD is substituted by imputed value $\hat{y}_{gs}$ .We call this gene expression matrix ID (imputed data).

Step 4: Perform prediction analysis using the imputed data (ID) and assess the Youden's index by YI=sensitivity+specificity-1.

Step5: Repeat step2-4 for *N* times. We use *N*=100 in this thesis.

2.5.3 Adjusted Rand Index (ARI) for gene clustering analysis

Adjust Rand index (ARI) by (Hubert, 2001) is commonly used to evaluate similarity of 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 pairs. The adjusted Rand index (ARI) 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 the data, we pretend that the cluster result from complete data is the gold standard. The clustering result from each imputed data is then compared to the gold standard by ARI. A higher ARI value indicates that the two clustering results are similar and the missing value imputation brings smaller impact to down-stream clustering analysis.

Step1: Given a real gene expression matrix, the (real) MVs are removed to form a complete gene expression matrix without MVs. We call this gene expression matrix CD (complete data). It is denoted by CD=$(y_{gs})_{G\times S}$

Step2: Next, a proportion q, 0<q<1, of MVs are intentionally introduced by randomly removing values in CD, where q=0.01, 0.05, 0.1 and 0.2. Let us denote this gene expression matrix with MD (missing data).

Step3: Imputation methods are applied to estimate MVs in MD. For missing entries, missing value, $(y_{gs})_{G\times S}$in MD is substituted by imputed value $\hat{y}_{gs}$ .We refer this to gene expression matrix ID (imputed data).

Step 4: Given a gene clustering method (k-means, SOM or Mclust) and the number of clusters (*K*),we can obtain gene clustering from complete data (C_{CD}) and from imputed data (C_{ID}). The ARI value is then computed, ARI(C_{CD}, C_{ID}). Since we do not know the number of clusters in each microarray data set, we perform evaluation for *K*=5, 10 and 15 and pick the *K* that has the highest ARI.

Step 5: Repeat 100 times from Step 2 to Step3.

3.0 RESULTS

3.1 COMPARISON OF CONSISTENCY MEASURES AMONG RMSE MEASURES AND AMONG DOWN-STREAM ANALYSIS METHODS

To answer Aim 1A, Table 3A and 3B shows the median consistency measure for RMSE measures, $\tilde{r}^{RMSE\times RMSE}$, in the eight data sets for DE gene detection and classification and three data sets for gene clustering. It is easily seen that as the missing value percentage increases (from 1€% to 20%), the consistency measure for selecting MV imputation method between RMSE measure also significantly increases. In some data sets, the consistency measure of NRMSE vs LRMSE is the highest while, in some other data sets, RAE vs NRMSE has the highest consistency measure. At the largest missing percentage, NRMSE vs LRMSE also have the highest consistency measure in ALO, SIN, VAN, YU, and BEE. From the variable and often low to intermediate consistency measures, we conclude that the decisions made by NRMSE, LRMSE and RAE for evaluating MV imputation methods are often quite different.

For Aim 1B, Table 4A, 4B and 4C are generated to compare consistency measures ($\tilde{r}^{BLCI\times BLCI}$, $\tilde{r}^{YI\times YI}$ and $\tilde{r}^{ARI\times ARI}$) of different down-stream analysis methods. Similar to Table 3A and 3B, the consistency measures are also significantly increased as the missing value p€ercentages€ increase from 1% to 20%. In Table 4A for DE gene detection, the consistency measures between the three methods (SAM, LIMMA and t-test+BH) are very high (around 0.5-0.9 in 20% missing value percentage). For classification, in Table 4B, the consistency measures are relatively lower (around 0.1~0.45 in 20% missing value percentage), indicating less impact of missing value imputation on these two down-stream analysis. Note that there are some NA values because if the largest Youden's Indices for 8 imputation methods are exactly same, then we are unable to compute correlation values. For gene clustering in Table 4C, consistency measures for K-means (k=5) and SOM (k=5, 10, and 15) present a good consistency. Thus, in gene clustering, the selection of K value is more important prior to clustering analysis and estimation of MV.

3.2 WHICH RMSE MEASURE BETTER CORRELATES WITH BIOLOGICAL IMPACT MEASURES?

Table 5-A, 5-B and 5-C show the consistency measures between RMSE measures and down-stream analysis methods ($\tilde{r}^{RMSE\times BLCI}$, $\tilde{r}^{RMSE\times YI}$ and $\tilde{r}^{RMSE\times ARI}$) to answer Aim 2: "which RMSE measure is better correlated with biological impact measures?". In Table 5A, the consistency measures for LRMSE with all three DE gene detection methods (SAM, LIMMA and t-test+BH) are clearly the highest, followed by NRMSE and RAE is the lowest. Again, the consistency measures are increased as missing value percentage increase. For the 20% missing value percentage, the consistency measure of LRMSE and BLCI from SAM, LIMMA, and t+BH are as high as 0.6- 0.9. In Table 5-B, the consistency measures even for 20% missing value percentage are much lower than 5A (around -0.2-0.5) while LRMSE still clearly outperforms RAE and followed by NRMSE. The low consistency measures suggest that classification is hardly affected missing value imputation. In Table 5-C, the consistency measure for 20% missing value are very well correlated with LRMSE in K-means (k=5) and SOM (k=5, 10, and 15). Table 5A-5C provides an effective way to determine that LRMSE is a better quantitative measure than RAE and NRMSE to correlate with biological measures in ranking performance of

MV imputation methods. We further apply a complementary approach by linear regression model to better quantify the degree of consistency. The estimate ($\beta^{*(BLCI)}$; the median slope estimate of 100 simulations) and statistical significance ($(\beta^{L;(BLCI)},\beta^{H;(BLCI)})$; the 95% confidence interval of the slope estimates in 100 simulations) of the slope term of the linear model indicate the ability to predict biological impacts by RMSE measures in different MV imputation methods. Since we have concluded that LRMSE is more consistent with the biological impacts, we only perform the linear models for LRMSE in Table 6A for DE gene detection, 6B for classification and 6C for gene clustering. In Table 6A, we can clearly see that the slope estimates are negative in almost all situations and the slope decreases when MV percentage increases. At 1% MV percentage, At 20% MV percentage, all data sets are negative with statistical significance (i.e. the 95% confidence intervals do not cover zero). Except VAN and BEE, the R^2 values are as high as around 0.45-0.86 and corresponding coefficient of slope is also significant. For classification and clustering in Table 6B, the statistical significance and R^2 are much weaker, indicating data imputation method does not affect much the performances of the classifiers. Especially, NA for some of consistency measures represents its evidence, which all Youden Indices corresponding to 8 imputation methods are exactly same, then correlation value is not able to be computed, that is, returns NA value. Thus, estimating MV does not improve much the classification accuracy and since consistency measures between RMSE vs Youden's Index in classification does not present a good relationship and the ranking of MV methods by RMSE based measure does not guarantee the same order of raking by biological impact measure such as Youden's Index, the choice of MV methods should be determined in context of classification accuracy for disease classifiers.

3.3 TABLES

Table 3-A.Consistency measure among 3 RMSEs in DE.

			GOL			ALO			LUO			SIN	
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	NRMSE	1	0.833	0.190	1	0.690	0.690	1	0.214	0.619	1	0.286	0.429
	LRMSE		1	0.143		1	0.405		1	0.440		1	0.25
	RAE			1			1			1			1
5%	NRMSE	1	0.929	0.298	1	0.667	0.655	1	0.476	0.643	1	0.310	0.381
	LRMSE		1	0.190		1	0.405		1	0.464		1	0.10
	RAE			1			1			1			1
10%	NRMSE	1	0.833	0.405	1	0.702	0.667	1	0.571	0.762	1	0.119	0.262
	LRMSE		1	0.024		1	0.429		1	0.429		1	0.262
	RAE			1			1			1			1
20%	NRMSE	1	0.929	0.524	1	0.845	0.714	1	0.619	0.881	1	0.452	0.774
	LRMSE		1	0.476		1	0.738		1	0.429		1	0.643
	RAE			1			1			1			1
			LAP			VAN			YU			BEE	
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	NRMSE	1	-0.083	0.714	1	0.167	0.190	1	0.762	0.833	1	0.690	0.690
	LRMSE		1	0.119		1	0.571		1	0.714		1	0.405
	RAE			1			1			1			1
5%	NRMSE	1	0.000	0.079	1	0.369	0.143	1	0.762	0.762	1	0.667	0.655
	LRMSE		1	0.190		1	0.548		1	0.595		1	0.405
	RAE			1			1			1			1
10%	NRMSE	1	0.095	0.738	1	0.310	0.333	1	0.833	0.762	1	0.702	0.667
	LRMSE		1	0.357		1	0.548		1	0.619		1	0.429
	RAE			1			1			1			1
20%	NRMSE	1	0.429	0.810	1	0.524	0.357	1	0.857	0.905	1	0.845	0.714
	LRMSE		1	0.595		1	0.619		1	0.690		1	0.738
	RAE			1			1			1			1

It represents the spearman rank correlation values among 3 different RMSE measures for DE datasets.

Table 3-B. Consistency measure among 3 RMSEs in gene clustering.

			SP.AFA			SP.ELU			CAU	
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
	NRMSE	1	-0.024	0.905	1	0.143	0.929	1	0.631	0.286
1%	LRMSE		1	0.095		1	0.155		1	-0.476
	RAE			1			1			1
	NRMSE	1	-0.071	0.881	1	0.524	0.952	1	0.667	0.286
5%	LRMSE		1	-0.036		1	0.524		1	-0.452
	RAE			1			1			1
	NRMSE	1	-0.119	0.774	1	0.143	0.952	1	0.357	0.286
10%	LRMSE		1	-0.071		1	0.119		1	-0.667
	RAE			1			1			1
	NRMSE	1	0.036	0.726	1	0.381	0.976	1	0.762	0.333
20%	LRMSE		1	-0.167		1	0.381		1	-0.071
	RAE			1			1			1

It represents the spearman rank correlation values among 3 different RMSE measures for clustering datasets.

Table 4-A. Consistency measure among 3 DE methods.

			GOL			ALO			LUO			SIN	
		SAM	LIMMA	t-BH	SAM	LIMMA	t-BH	SAM	LIMMA	t-BH	SAM	LIMMA	t-BH
	SAM	1	0.084	0.114	1	0.069	0.031	1	-0.048	0.096	1	-0.078	0.048
1%	LIMMA		1	0.560		1	0.380		1		0.383	1	0.238
	t-BH			1			1			1			1
	SAM	1	0.286	0.286	1	0.371	0.310	1	0.452	0.405	1	0.027	-0.090
5%	LIMMA		1	0.820		1	0.788		1	0.774		1	0.548
	t-BH			1			1			1			1
	SAM	1	0.464	0.554	1	0.524	0.531	1	0.571	0.571	1	0.083	0.179
10%	LIMMA		1	0.905		1	0.905		1	0.881		1	0.786
	t-BH			1			1			1			1
	SAM	1	0.702	0.714	1	0.571	0.571	1	0.762	0.762	1	0.476	0.476
20%	LIMMA		1	0.952		1	0.929		1	0.929		1	0.929

Table 4-A continued.

	t-BH			1			1			1			1
			LAP			VAN			YU			BEE	
1%	SAM	1	0.006	0.024	1	0.054	0.083	1	0.024	0.120	1	0.272	0.286
	LIMMA		1	0.383		1	0.366		1	0.641		1	0.714
	t-BH			1			1			1			1
5%	SAM	1	0.095	0.214	1	0.083	0.048	1	0.476	0.533	1	0.524	0.524
	LIMMA		1	0.758		1	0.653		1	0.844		1	0.929
	t-BH			1			1			1			1
10%	SAM	1	0.298	0.262	1	0.321	0.251	1	0.595	0.643	1	0.667	0.667
	LIMMA		1	0.833		1	0.738		1	0.905		1	0.929
	t-BH			1			1			1			1
20%	SAM	1	0.667	0.667	1	0.714	0.690	1	0.810	0.833	1	0.929	0.929
	LIMMA		1	0.952		1	0.913		1	0.976		1	0.976
	t-BH			1			1			1			1

It represents the spearman rank correlation values among 3 DE methods using 8 datasets.

Table 4-B. Consistency measure among 4 classification methods.

			GOL				ALO				LUO		
		LDA	KNN	SVM	PAM	LDA	KNN	SVM	PAM	LDA	KNN	SVM	PAM
1%	LDA	1	0.181	-0.218	-0.370	1	-0.080	0.000	0.132	NA	NA	NA	NA
	KNN		1	-0.302	-0.003		1	0.040	0.000	NA	1	NA	NA
	SVM			1	-0.286			1	-0.108	NA	NA	NA	NA
	PAM				1				1	NA	NA	NA	NA
5%	LDA	1	0.246	0.038	0.146	1	0.059	0.187	-0.052	1	0.655	NA	NA
	KNN		1	0.271	0.062		1	0.128	0.293		1	NA	NA
	SVM			1	0.143			1	0.058			1	NA
	PAM				1				1				1
10%	LDA	1	0.351	0.029	0.176	1	0.057	0.226	0.170	1	0.571	NA	NA
	KNN		1	0.269	0.095		1	-0.013	0.218		1	NA	NA
	SVM			1	0.143			1	0.113			1	NA

Table 4-B continued.

	PAM				1				1				1
20%	LDA	1	0.249	0.200	0.255	1	0.203	0.448	0.165	1	0.381	0.256	NA
	KNN		1	0.187	0.255		1	0.150	0.382		1	0.488	NA
	SVM			1	0.143			1	0.107			1	NA
	PAM				1				1				1
			SIN				LAP				VAN		
		LDA	KNN	SVM	PAM	LDA	KNN	SVM	PAM	LDA	KNN	SVM	PAM
1%	LDA	1	0.075	0.616	-0.325	1	-0.014	-0.029	-0.041	1	0.103	0.168	0.046
	KNN		1	0.361	0.143		1	-0.057	0.143		1	-0.018	0.000
	SVM			1	0.285			1	0.143			1	0.000
	PAM				1				1				1
5%	LDA	1	0.248	0.403	-0.071	1	0.085	0.114	0.173	1	0.051	0.262	0.172
	KNN		1	0.160	0.262		1	0.166	0.293		1	0.100	0.048
	SVM			1	-0.014			1	0.170			1	0.113
	PAM				1				1				1
10%	LDA	1	0.348	0.344	0.132	1	0.000	0.177	0.037	1	0.090	0.218	0.486
	KNN		1	0.212	0.320		1	0.216	0.351		1	0.140	0.123
	SVM			1	0.163			1	0			1	0.192
	PAM				1				1				1
20%	LDA	1	0.483	0.647	0.500	1	0.319	0.440	0.311	1	0.248	0.231	0.619
	KNN		1	0.506	0.434		1	0.477	0.531		1	0.210	0.126
	SVM			1	0.287			1	0.423			1	0.187
	PAM				1				1				1
			YU				BEE						
		LDA	KNN	SVM	PAM	LDA	KNN	SVM	PAM				
1%	LDA	1	0.000	-0.037	-0.072	1	0.271	-0.143	NA				
	KNN		1	0.000	0.105		1	0.218	NA				
	SVM			1	-0.143			1	0.000				
	PAM				1				1				
5%	LDA	1	0.000	0.000	0.115	1	0.000	0.218	0.060				
	KNN		1	0.066	-0.103		1	-0.255	0.281				
	SVM			1	0.000			1	0.258				

Table 4-B continued.

	PAM				1				1
10%	LDA	1	0.000	0.010	0.158	1	0.000	0.293	0.333
	KNN		1	-0.153	-0.014		1	0.000	0.390
	SVM			1	-0.073			1	0.238
	PAM				1				1
20%	LDA	1	0.345	0.230	0.132	1	0.138	0.119	0.293
	KNN		1	0.364	0.029		1	0.218	0.293
	SVM			1	-0.024			1	0.282
	PAM				1				1

It represents the spearman rank correlation values among 4 classification methods using 8 datasets.

Table 4-C. Consistency measure between 2 clustering methods.

SP.AFA			Kmeans			SOM		
			K=5	K=10	K=15	K=5	K=10	K=15
1%	Kmeans	K=5	1.000	-0.071	0.060	0.238	0.167	0.333
		K=10		1	0.107	0.060	0.012	0.000
		K=15			1	0.024	0.036	0.048
	SOM	K=5				1	0.036	0.738
		K=10					1	0.119
		K=15						1
5%	Kmeans	K=5	1.000	0.143	0.131	0.417	0.452	0.643
		K=10		1	-0.024	0.036	0.095	0.190
		K=15			1	0.071	0.071	0.071
	SOM	K=5				1	0.238	0.547
		K=10					1	0.643
		K=15						1
10%	Kmeans	K=5	1.000	0.131	0.214	0.429	0.702	0.786
		K=10		1	0.143	0.143	0.167	0.190
		K=15			1	0.167	0.143	0.190
	SOM	K=5				1	0.286	0.405
		K=10					1	0.810

Table 4-C continued.

		K=15						1
20%	Kmeans	K=5	1.000	0.369	0.393	0.381	0.738	0.786
		K=10		1	0.262	0.190	0.357	0.345
		K=15			1	0.036	0.381	0.429
	SOM	K=5				1	0.357	0.381
		K=10					1	0.905
		K=15						1
SP.ELU			Kmeans			SOM		
			K=5	K=10	K=15	K=5	K=10	K=15
1%	Kmeans	K=5	1.000	0.024	0.071	0.286	0.286	0.262
		K=10		1	0.024	0.024	0.048	0.000
		K=15			1	0.071	0.095	0.119
	SOM	K=5				1	0.833	0.857
		K=10					1	0.095
		K=15						1
5%	Kmeans	K=5	1.000	0.060	0.024	0.321	0.333	0.405
		K=10		1	0.083	0.119	0.107	0.083
		K=15			1	0.107	0.143	0.190
	SOM	K=5				1	0.857	0.833
		K=10					1	0.929
		K=15						1
10%	Kmeans	K=5	1.000	0.024	0.095	0.452	0.417	0.440
		K=10		1	0.167	-0.048	-0.036	-0.024
		K=15			1	0.155	0.143	0.155
	SOM	K=5				1	0.952	0.929
		K=10					1	0.976
		K=15						1
20%	Kmeans	K=5	1.000	0.476	0.488	0.833	0.833	0.833
		K=10		1	0.595	0.524	0.512	0.524
		K=15			1	0.524	0.524	0.524
	SOM	K=5				1	0.976	0.976
		K=10					1	0.976

Table 4-C continued.

		K=15						1

CAU			Kmeans			SOM		
			K=5	K=10	K=15	K=5	K=10	K=15
1%	Kmeans	K=5	1.000	0.060	0.000	0.143	0.113	0.238
		K=10		1	0.036	0.096	0.012	0.000
		K=15			1	-0.071	-0.107	-0.107
	SOM	K=5				1	0.476	0.524
		K=10					1	0.583
		K=15						1
5%	Kmeans	K=5	1.000	-0.024	0.000	0.202	0.179	0.262
		K=10		1	0.024	0.071	0.107	0.036
		K=15			1	-0.024	0.000	-0.012
	SOM	K=5				1	0.810	0.75
		K=10					1	0.738
		K=15						1
10%	Kmeans	K=5	1.000	0.048	0.048	0.429	0.333	0.345
		K=10		1	0.000	0.048	0.048	0.048
		K=15			1	0.048	-0.024	0.071
	SOM	K=5				1	0.786	0.726
		K=10					1	0.571
		K=15						1
20%	Kmeans	K=5	1.000	0.321	0.452	0.619	0.476	0.393
		K=10		1	0.310	0.357	0.333	0.333
		K=15			1	0.452	0.452	0.345
	SOM	K=5				1	0.726	0.619
		K=10					1	0.476
		K=15						1

It represents the spearman rank correlation values among2 different clustering methods with 3 different cluster numbers, k=5,10, and 15 using 3 datasets.

Table 5-A. Consistency measure between RMSEs and DE.

		GOL			ALO			LUO			SIN		
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	SAM	0.305	**0.353**	-0.030	**0.179**	0.119	0.143	0.000	**0.155**	0.036	-0.024	**0.048**	0.024
	LIMMA	0.313	**0.440**	0.024	0.453	**0.578**	0.389	0.095	**0.353**	0.204	0.060	**0.381**	-0.048
	t-BH	0.296	**0.371**	-0.180	0.375	**0.524**	0.096	0.072	**0.310**	0.114	0.071	**0.257**	0.060
5%	SAM	0.429	**0.496**	-0.071	0.381	**0.464**	0.262	0.036	**0.262**	0.000	-0.012	**0.083**	0.071
	LIMMA	0.460	**0.631**	-0.305	0.513	**0.710**	0.095	0.012	**0.381**	-0.167	0.048	**0.619**	-0.179
	t-BH	0.452	**0.595**	-0.394	0.543	**0.716**	0.113	0.048	**0.393**	-0.143	0.071	**0.597**	-0.083
10%	SAM	0.531	**0.615**	-0.060	0.548	**0.643**	0.333	0.024	**0.405**	-0.071	-0.048	**0.167**	0.060
	LIMMA	0.381	**0.647**	-0.333	0.429	**0.719**	-0.048	0.024	**0.452**	-0.202	-0.131	**0.567**	-0.179
	t-BH	0.281	**0.619**	-0.500	0.429	**0.690**	-0.060	-0.048	**0.405**	-0.214	-0.095	**0.667**	-0.131
20%	SAM	0.762	**0.798**	0.316	0.619	**0.714**	0.405	0.143	**0.548**	-0.048	0.048	**0.524**	0.202
	LIMMA	0.561	**0.643**	-0.071	0.262	**0.405**	-0.083	-0.012	**0.548**	-0.179	-0.191	**0.619**	0.024
	t-BH	0.512	**0.619**	-0.131	0.262	**0.383**	-0.012	-0.048	**0.500**	-0.214	-0.119	**0.667**	0.048
		LAP			VAN			YU			BEE		
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	SAM	0.024	**0.048**	0.048	**0.048**	0.000	-0.048	0.048	**0.143**	0.048	-0.048	**0.155**	-0.310
	LIMMA	-0.189	**0.335**	-0.156	0.023	**0.224**	0.096	0.497	**0.683**	0.429	-0.012	**0.557**	-0.500
	t-BH	-0.200	**0.265**	-0.181	-0.055	**0.222**	0.060	0.527	**0.683**	0.393	-0.018	**0.444**	-0.466
5%	SAM	-0.036	**0.179**	0.048	-0.095	**0.071**	0.036	0.381	**0.619**	0.214	-0.024	**0.405**	-0.357
	LIMMA	-0.452	**0.488**	-0.333	-0.084	**0.257**	0.084	0.655	**0.833**	0.452	-0.071	**0.619**	-0.603
	t-BH	-0.449	**0.476**	-0.333	-0.048	**0.321**	0.167	0.548	**0.810**	0.381	-0.071	**0.619**	-0.595
10%	SAM	-0.167	**0.321**	-0.119	-0.190	**0.143**	0.000	0.524	**0.690**	0.250	-0.095	**0.405**	-0.405
	LIMMA	-0.512	**0.464**	-0.417	-0.167	**0.173**	-0.274	0.714	**0.810**	0.429	-0.119	**0.548**	-0.619
	t-BH	-0.452	**0.500**	-0.328	-0.155	**0.262**	-0.143	0.655	**0.806**	0.333	-0.119	**0.571**	-0.607
20%	SAM	0.048	**0.667**	0.167	0.238	**0.512**	0.190	0.667	**0.833**	0.476	0.036	**0.619**	-0.238
	LIMMA	-0.262	**0.429**	-0.143	0.214	**0.472**	0.119	0.786	**0.929**	0.619	0.095	**0.667**	-0.190
	t-BH	-0.286	**0.429**	-0.143	0.190	**0.476**	0.119	0.786	**0.929**	0.595	0.095	**0.667**	-0.167

It represents the spearman rank correlation values between 3 RMSE measures and BLCI after 3 DE methods using 8 datasets.

Table 5-B. Consistency measure between RMSEs and Classification.

		GOL			ALO			LUO		
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	LDA	**0.000**	-0.065	-0.169	**0.119**	0.013	-0.014	NA	**NA**	NA
	KNN	0.323	**0.395**	-0.074	-0.019	**0.052**	-0.052	-0.577	**0.247**	-0.247
	SVM	0.083	**0.577**	0.394	**-0.126**	-0.169	-0.126	NA	**NA**	NA
	PAM	0.412	**0.412**	-0.096	**-0.169**	-0.252	-0.218	NA	**NA**	NA
5%	LDA	0.051	**0.051**	0.103	**0.176**	0.145	0.082	**0.167**	-0.124	-0.332
	KNN	0.000	**0.126**	0.000	-0.167	**-0.124**	-0.160	**0.041**	0.000	-0.458
	SVM	-0.247	**0.000**	-0.083	-0.192	-0.154	**-0.096**	NA	**NA**	NA
	PAM	0.394	**0.412**	-0.169	-0.080	**0.176**	-0.327	NA	**NA**	NA
10%	LDA	0.038	**0.109**	-0.083	0.157	**0.170**	0.098	-0.041	**0.148**	-0.412
	KNN	0.000	**0.104**	-0.083	-0.262	-0.254	**-0.244**	-0.211	**0.028**	-0.738
	SVM	0.006	**0.083**	0.116	-0.149	-0.105	**0.000**	NA	**NA**	NA
	PAM	0.169	**0.252**	-0.247	-0.083	**0.196**	-0.412	NA	**NA**	NA
20%	LDA	0.056	**0.109**	0.095	0.246	**0.280**	0.175	-0.247	**0.118**	-0.412
	KNN	0.094	**0.096**	0.000	-0.230	**-0.209**	-0.221	-0.378	**0.056**	-0.577
	SVM	0.100	**0.109**	0.056	-0.038	**0.038**	-0.171	-0.412	**-0.247**	-0.577
	PAM	0.247	**0.247**	-0.100	-0.050	**0.167**	-0.183	NA	**NA**	NA
		SIN			LAP			VAN		
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	LDA	**-0.083**	-0.267	-0.104	0.039	0.036	0.073	-0.030	**0.073**	-0.037
	KNN	0.000	**0.041**	-0.083	0.083	-0.208	0.083	0.043	0.025	**0.048**
	SVM	0.096	**0.109**	-0.137	-0.149	-0.069	-0.167	0.030	**0.066**	-0.024
	PAM	0.126	**0.169**	-0.027	-0.083	-0.250	-0.218	**0.024**	-0.083	-0.083
5%	LDA	-0.103	**0.000**	-0.086	**-0.096**	-0.108	-0.144	**-0.048**	-0.067	-0.073
	KNN	0.077	**0.232**	-0.154	-0.282	**0.069**	-0.297	**0.018**	-0.090	-0.108
	SVM	0.000	**0.083**	-0.094	-0.254	**0.188**	-0.313	0.000	**0.024**	-0.031
	PAM	-0.028	**0.287**	-0.109	-0.126	**0.041**	-0.268	**-0.096**	-0.144	-0.145
10%	LDA	0.000	**0.261**	-0.052	-0.060	**-0.012**	-0.090	-0.078	-0.072	**0.036**
	KNN	0.056	**0.279**	-0.218	-0.252	**0.252**	-0.252	-0.095	-0.108	**-0.030**
	SVM	0.000	**0.148**	-0.083	-0.276	**0.250**	-0.313	-0.120	**0.006**	0.024
	PAM	-0.061	**0.257**	-0.200	-0.281	**0.089**	-0.252	-0.096	-0.048	-0.119

Table 5-B continued.

20%	LDA	-0.006	**0.503**	0.102	-0.246	**0.096**	-0.227	0.036	**0.060**	0.024
	KNN	-0.066	**0.486**	0.037	-0.401	**0.126**	-0.295	0.000	**0.107**	0.012
	SVM	-0.132	**0.390**	0.048	-0.294	**0.194**	-0.210	0.018	**0.084**	0.048
	PAM	0.062	**0.479**	0.106	-0.482	**-0.014**	-0.412	0.012	**0.107**	0.024

			YU			BEE	
		NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	LDA	**0.013**	0.000	0.098	-0.083	-0.083	-0.126
	KNN	-0.059	**0.174**	0.080	**0.056**	0.000	-0.218
	SVM	0.247	**0.394**	0.309	0.028	-0.323	**0.071**
	PAM	0.000	**0.078**	-0.078	0.225	-0.507	**0.394**
5%	LDA	0.027	**0.050**	-0.052	-0.126	**-0.083**	-0.247
	KNN	**0.239**	0.218	0.163	-0.083	**-0.013**	0.028
	SVM	0.157	**0.275**	0.091	-0.083	0.096	-0.126
	PAM	0.013	**0.145**	-0.213	-0.056	-0.247	**0.252**
10%	LDA	-0.047	**0.062**	-0.150	-0.056	**0.083**	-0.252
	KNN	**0.325**	0.321	0.288	-0.083	-0.169	**-0.083**
	SVM	0.114	**0.185**	0.025	-0.109	**0.083**	-0.378
	PAM	-0.145	**0.135**	-0.349	**0.069**	-0.109	0.000
20%	LDA	0.134	**0.314**	0.062	-0.083	**-0.006**	-0.126
	KNN	0.457	**0.481**	0.426	-0.136	-0.247	**-0.078**
	SVM	-0.169	**0.357**	0.109	-0.247	**-0.083**	-0.169
	PAM	**0.224**	0.132	-0.342	-0.378	**-0.247**	-0.412

It represents the spearman rank correlation values between 3 RMSE measures and Youden’s Index after 4 classification methods using 8 datasets.

Table 5-C. Consistency measure between RMSEs and Clustering.

			SP.AFA			SP.ELU			CAU		
			NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE	NRMSE	LRMSE	RAE
1%	Kmeans	K=5	-0.024	**0.333**	0.333	-0.036	**0.286**	-0.048	0.048	**0.274**	-0.167
		K=10	-0.060	**-0.012**	-0.012	0.036	**-0.060**	0.071	0.071	**0.071**	-0.024
		K=15	0.071	**0.107**	0.107	-0.036	**0.024**	0.012	-0.083	**0.036**	-0.024
	SOM	K=5	-0.119	**0.798**	0.000	-0.048	**0.833**	0.000	0.5	**0.690**	-0.167
		K=10	-0.214	**0.095**	-0.226	0.024	**0.893**	0.071	0.488	**0.722**	-0.143
		K=15	-0.095	**0.119**	-0.024	0.024	**0.905**	0.095	0.488	**0.583**	-0.298
5%	Kmeans	K=5	0.000	**0.690**	0.048	0.226	**0.405**	0.226	0.190	**0.238**	-0.083
		K=10	0.024	**0.226**	0.024	0.119	**0.167**	0.139	0.071	**0.095**	-0.036
		K=15	0.024	**-0.024**	-0.012	0.048	**0.083**	0.095	-0.071	**0.024**	-0.048
	SOM	K=5	0.071	**0.619**	0.119	0.300	**0.833**	0.357	0.690	**0.857**	-0.298
		K=10	-0.262	**0.619**	-0.381	0.440	**0.881**	0.524	0.643	**0.881**	-0.321
		K=15	-0.143	**0.643**	-0.167	0.429	**0.929**	0.5	0.631	**0.738**	-0.333
10%	Kmeans	K=5	0.024	**0.810**	0.012	0.048	**0.452**	0.024	0.107	**0.429**	-0.333
		K=10	0.024	**0.190**	0.071	0.143	-0.048	**0.167**	**0.048**	0.036	0.036
		K=15	-0.012	**0.143**	-0.083	0.119	**0.167**	0.143	**0.048**	0.000	-0.095
	SOM	K=5	0.143	**0.429**	0.214	0.000	**0.929**	-0.024	0.310	**0.905**	-0.619
		K=10	-0.214	**0.833**	-0.274	0.095	**0.929**	0.095	0.310	**0.833**	-0.548
		K=15	-0.143	**0.810**	-0.190	0.143	**0.976**	0.143	0.214	**0.571**	-0.488
20%	Kmeans	K=5	0.083	**0.690**	0.095	0.286	**0.857**	0.333	0.548	**0.607**	0.143
		K=10	-0.012	**0.333**	0.071	0.440	**0.548**	0.440	0.333	**0.357**	0.226
		K=15	0.000	**0.262**	0.024	0.524	**0.595**	0.619	**0.464**	0.310	0.226
	SOM	K=5	0.119	**0.464**	0.226	0.286	**0.976**	0.333	0.810	**0.881**	0.000
		K=10	-0.036	**0.905**	-0.214	0.286	**0.976**	0.357	0.583	**0.702**	0.000
		K=15	-0.048	**0.905**	-0.155	0.286	**0.976**	0.333	0.560	**0.476**	0.095

It represents the spearman rank correlation values between 3 RMSE measures and ARI after 2 clustering methods for different cluster number, k=5,10,and 15 using 3 datasets.

Table 6-A. Linear Model between LRMSE and DE.

		GOL	ALO	LUO	SIN
		β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2
1%	SAM	-0.039(-0.408,0.365);0.072	-0.002(-0.234,0.176);0.063	-0.016(-1.176,1.021);0.073	-0.036(-1.354,1.242);0.071
	LIMMA	-0.031(-0.119,0.056);0.234	-0.072(-0.192,0.007);0.619	**-0.045*(-0.106,-0.001);0.540**	-0.037(-0.190,0.068);0.199
	t-BH	-0.032(-0.120,0.042);0.162	-0.099(-0.315,0.045);0.529	-0.0439(-0.107,0.007);0.413	-0.035(-0.137,0.058);0.131
5%	SAM	-0.010(-0.420,0.300);0.086	-0.106(-0.321,0.063);0.195	-0.105(-0.865,0.333);0.099	-0.131(-1.372,1.063);0.110
	LIMMA	-0.129(-0.324,0.025);0.355	**-0.410*(-0.627,-0.232);0.737**	**-0.208*(-0.323,-0.108);0.714**	-0.166(-0.359,0.041);0.431
	t-BH	-0.151(-0.314,0.042);0.308	**-0.509*(-1.031,-0.267);0.732**	**-0.210*(-0.332,-0.107);0.632**	-0.157(-0.362,0.045);0.384
10%	SAM	-0.193(-0.594,0.134);0.146	**-0.229*(-0.597,-0.054);0.661**	-0.301(-1.011,0.012);0.341	-0.168(-1.017,0.842);0.094
	LIMMA	**-0.253*(-0.477,-0.078);0.348**	**-0.780*(-1.168,-0.491);0.768**	**-0.353*(-0.519,-0.212);0.721**	**-0.245(-0.430,-0.028);0.440**
	t-BH	**-0.272*(-0.541,-0.106);0.310**	**-0.976*(-1.466,-0.558);0.720**	**-0.356*(-0.499,-0.221);0.661**	**-0.263(-0.474,-0.055);0.512**
20%	SAM	**-0.213*(-0.337,-0.075);0.399**	**-0.353*(-0.695,-0.155);0.636**	**-0.399*(-0.701,-0.197);0.676**	**-0.231(-0.524,0.048);0.328**
	LIMMA	**-0.256*(-0.358,-0.097);0.368**	**-1.111*(-1.601,-0.653);0.454**	**-0.502*(-0.700,-0.340);0.685**	**-0.314(-0.445,-0.177);0.669**
	t-BH	**-0.243*(-0.352,-0.121);0.279**	**-1.22*(-1.74,-0.69);0.390**	**-0.520*(-0.696,-0.334);0.636**	**-0.324(-0.450,-0.189);0.686**
		LAP	VAN	YU	BEE
		β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2
1%	SAM	-0.015(-0.625,0.693);0.061	-0.067(-3.488,2.554);0.110	-0.031(-0.138,0.168);0.090	-0.037(-1.519,1.282);0.090
	LIMMA	-0.062(-0.333,0.147);0.176	-0.218(-1.847,0.937);0.120	-0.028(-0.062,0.004);0.482	**-0.043*(-0.091,-0.001);0.310**
	t-BH	-0.019(-0.226,0.131);0.185	-0.231(-1.772,0.993);0.108	-0.028(-0.062,0.012);0.494	-0.449(-0.111,0.002);0.244
5%	SAM	-0.072(-1.336,1.517);0.087	-0.083(-2.839,2.708);0.096	-0.089(-0.248,0.064);0.350	-0.188(-1.048,0.734);0.167
	LIMMA	-0.191(-0.564,0.038);0.259	-0.539(-2.116,0.948);0.088	**-0.088*(-0.143,-0.016);0.695**	**-0.165*(-0.262,-0.089);0.290**
	t-BH	-0.219(-0.548,0.058);0.333	-0.748(-2.284,1.137);0.162	**-0.092*(-0.144,-0.037);0.667**	**-0.176*(-0.290,-0.086);0.270**
10%	SAM	-0.175(-0.710,0.283);0.131	-0.349(-2.386,1.877);0.091	**-0.147*(-0.237,-0.048);0.530**	**-0.274*(-0.897,0.317);0.125**
	LIMMA	**-0.401*(-0.788,-0.046);0.301**	-0.44(-2.3,1.18);0.050	**-0.121*(-0.209,-0.052);0.733**	**-0.271*(-0.382,-0.155);0.212**
	t-BH	**-0.422*(-0.786,-0.057);0.343**	-0.797(-2.407,1.003);0.080	**-0.124*(-0.213,-0.069);0.691**	**-0.280*(-0.425,-0.169);0.223**
20%	SAM	**-0.383*(-0.738,-0.209);0.764**	**-1.34(-2.09,-0.72);0.759**	**-0.182*(-0.270,-0.000);0.740**	**-0.268*(-0.429,-0.115);0.285**
	LIMMA	**-0.610*(-0.874,-0.339);0.560**	**-1.599(-2.427,-0.594);0.705**	**-0.177*(-0.230,-0.000);0.873**	**-0.259*(-0.334,-0.185);0.440**
	t-BH	**-0.606*(-0.840,-0.377);0.572**	**-1.681(-2.473,-0.748);0.748**	**-0.178*(-0.232,-0.000);0.867**	**-0.263*(-0.361,-0.191);0.445**

It represents the estimate of slope, its 95% confidence interval, and R-squared value in simple linear regression between LRSME and BCLI after 3 DE methods.* indicates the statistical significance for the slope from 95 % confidence interval.

Table 6-B. Linear model between LRMSE and Classification.

		GOL	ALO	LUO	SIN
		β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2
1%	LDA	-0.000(-0.201,0.242);0.410	-0.000(-0.031,0.261);0.190	-0.000(-0.000,-0.000);0.457	-0.000(-0.118,0.230);0.516
	KNN	-0.000(-0.398,0.219);0.481	-0.000(-0.154,0.268);0.454	-0.000(-0.000,-0.000);0.455	-0.000(-0.359,0.189);0.541
	SVM	-0.000(-0.054,-0.000);0.503	-0.038(-0.509,0.277);0.140	-0.000(-0.000,-0.000);0.457	-0.000(-0.316,0.091);0.512
	PAM	-0.000(-0.133,-0.000);0.542	-0.000(-0.138,0.269);0.293	-0.000(-0.000,-0.000);0.457	-0.000(-0.276,0.295);0.456
5%	LDA	-0.000(-0.562,0.405);0.265	-0.048(-0.459,0.496);0.185	-0.000(-0.000,-0.000);0.471	-0.000(-0.429,0.368);0.312
	KNN	-0.000(-0.377,0.334);0.265	-0.088(-0.310,0.137);0.243	-0.000(-0.000,0.066);0.476	-0.084(-0.653,0.342);0.158
	SVM	-0.000(-0.315,0.219);0.500	-0.084(-0.523,0.352);0.199	-0.000(-0.000,-0.000);0.484	-0.000(-0.543,0.437);0.216
	PAM	-0.000(-0.248,0.123);0.512	-0.016(-0.157,0.127);0.085	-0.000(-0.000,-0.000);0.484	-0.045(-0.563,0.294);0.247
10%	LDA	-0.022(-0.399,0.508);0.109	-0.074(-0.672,0.346);0.173	-0.000(-0.596,0.128);0.496	-0.102(-0.557,0.311);0.118
	KNN	-0.018(-0.538,0.430);0.153	-0.041(-0.266,0.140);0.101	-0.000(-0.000,-0.000);0.509	-0.212(-0.762,0.240);0.137
	SVM	-0.000(-0.317,0.264);0.166	-0.090(-0.677,0.375);0.242	-0.000(-0.000,-0.000);0.512	-0.082(-0.653,0.448);0.131
	PAM	-0.000(-0.234,0.133);0.442	-0.033(-0.222,0.181);0.115	-0.000(-0.000,-0.000);0.512	-0.126(-0.504,0.206);0.154
20%	LDA	-0.025(-0.419,0.242);0.138	-0.191(-0.814,0.417);0.183	-0.000(-0.475,0.212);0.240	-0.190(-0.563,0.140);0.321
	KNN	-0.004(-0.226,0.228);0.081	-0.105(-0.480,0.165);0.159	-0.000(-0.000,-0.078);0.320	-0.176(-0.465,0.138);0.315
	SVM	-0.000(-0.231,0.268);0.138	-0.085(-0.591,0.336);0.099	-0.000(-0.000,0.046);0.321	-0.21(-0.57,0.21);0.270
	PAM	-0.001(-0.134,0.188);0.208	-0.067(-0.297,0.122);0.077	-0.000(-0.000,-0.000);0.324	-0.116(-0.375,0.115);0.346
		LAP	VAN	YU	BEE
		β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2	β_1(CI);R^2
1%	LDA	0.092(-1.336,1.517);0.144	-0.037(-1.425,1.527);0.084	-0.000(-0.358,0.317);0.194	-0.000(-0.124,0.091);0.230
	KNN	-0.000(-0.000,0.392);0.524	-0.090(-1.484,1.234);0.085	0.011(-0.314,0.368);0.226	-0.000(-0.813,0.731);0.465
	SVM	-0.000(-0.000,1.35);0.289	-0.043(-1.798,1.754);0.119	-0.000(-0.165,0.144);0.466	-0.000(-0.002,0.055);0.474
	PAM	-0.000(-0.866,1.08);0.283	-0.000(-0.493,0.679);0.124	-0.000(-0.303,0.277);0.217	-0.000(-0.000,-0.000);0.481
5%	LDA	0.216(-2.632,2.320);0.138	0.077(-1.814,2.033);0.087	-0.027(-0.482,0.387);0.098	-0.000(-0.112,0.093);0.193
	KNN	-0.000(-1.29,0.868);0.266	0.181(-2.273,2.140);0.126	0.010(-0.285,0.325);0.156	-0.000(-0.560,0.513);0.352
	SVM	-0.277(-1.859,1.526);0.109	0.014(-3.294,2.419);0.095	-0.027(-0.404,0.190);0.188	-0.000(-0.092,0.094);0.469
	PAM	-0.008(-1.666,1.694);0.160	0.163(-0.603,1.164);0.081	-0.017(-0.292,0.301);0.081	-0.000(-0.003,0.053);0.488
10%	LDA	0.066(-1.933,2.00);0.120	0.154(-2.577,2.579);0.073	-0.000(-0.773,0.456);0.105	-0.007(-0.473,0.095);0.152
	KNN	-0.307(-1.967,0.751);0.223	0.132(-2.276,2.449);0.065	-0.049(-0.438,0.382);0.100	-0.000(-0.739,0.734);0.196
	SVM	-0.429(-2.533,1.519);0.120	0.061(-2.893,3.411);0.083	-0.107(-0.555,0.253);0.201	-0.000(-0.091,0.092);0.424
	PAM	-0.320(-3.732,1.319);0.109	0.021(-1.187,1.636);0.068	-0.005(-0.423,0.304);0.044	-0.000(-0.038,0.039);0.484
20%	LDA	-0.137(-1.404,1.324);0.106	-0.112(-1.681,0.920);0.183	-0.071(-0.543,0.177);0.160	-0.000(-0.309,0.070);0.078

Table 6-B continued.

	KNN	-0.167(-0.916,0.449);0.121	-0.100(-0.919,0.768);0.108	-0.132(-0.519,0.168);0.209	0.014(-0.278,0.325);0.118
	SVM	-0.464(-1.251,0.560);0.176	-0.237(-1.535,0.734);0.114	-0.091(-0.472,0.206);0.166	-0.000(-0.022,0.186);0.129
	PAM	-0.012(-1.149,1.090);0.065	-0.140(-0.792,0.715);0.236	0.000(-0.196,0.213);0.039	-0.000(-0.167,0.070);0.393

It represents the estimate of slope, its 95% confidence interval, and R-squared value in simple linear regression between LRSME and Youden's Index after 4 classification methods.* indicates the statistical significance for the slope from 95 % confidence interval.

Table 6-C. Linear model between LRMSE and clustering interval.

			SP.AFA	SP.ELU	CAU
			β_1(CI);R^2	**β_1(CI);R^2**	**β_1(CI);R^2**
1%	Kmeans	K=5	-0.106(-1.420,1.359);0.210	-0.192(-0.950,0.500);0.078	-0.029(-0.524,0.320);0.081
		K=10	-0.013(-0.900,0.708);0.091	0.023(-0.525,0.720);0.145	-0.426(-2.821,2.631);0.081
		K=15	-0.020(-0.496,0.428);0.057	-0.047(-0.445,0.365);0.071	0.297(-2.160,2.351);0.092
	SOM	K=5	**-0.644*(-1.386,-0.190);0.712**	**-0.174*(-0.211,-0.135);0.960**	-0.022(-0.052,0.001);0.575
		K=10	-0.076(-0.257,0.061);0.145	**-0.107*(-0.147,-0.084);0.873**	**-0.024*(-0.049,-0.001);0.675**
		K=15	**-0.211*(-0.313,-0.144);0.916**	**-0.122*(-0.150,-0.086);0.982**	-0.028*(-2.916,0.002);0.673
5%	Kmeans	K=5	-0.572(-2.346,3.745);0.637	-0.001(-2.313,1.551);0.079	-0.073(-0.745,0.246);0.100
		K=10	-0.339(-2.467,1.842);0.145	-0.012(-2.112,0.953);0.108	-0.080(-2.512,2.005);0.080
		K=15	-0.307(-1.461,0.811);0.140	-0.087(-1.388,0.551);0.075	0.028(-0.975,1.014);0.101
	SOM	K=5	-2.007(-2.961,0.291);0.471	**-0.613*(-0.830,-0.000);0.863**	**-0.074(-0.116,-0.025);0.833**
		K=10	**-0.728*(-1.063,-0.374);0.768**	**-0.695*(-0.949,-0.000);0.926**	-0.087(-0.814,0.003);0.817
		K=15	**-1.080*(-1.341,-0.848);0.916**	**-0.615*(-0.720,-0.000);0.945**	-0.097(-2.959,1.297);0.863
10%	Kmeans	K=5	-1.031(-4.639,3.067);0.548	-0.000(-0.095,-0.000);0.170	-0.121(-0.667,0.284);0.156
		K=10	-0.819(-3.682,1.705);0.096	0.000(-0.001,0.074);0.194	-0.323(-2.766,2.082);0.068
		K=15	-0.346(-1.683,0.916);0.102	0.000(-0.001,0.032);0.079	0.052(-0.867,0.829);0.081
	SOM	K=5	**-1.613*(-3.509,-0.027);0.289**	-0.000(-0.139,-0.000);0.966	**-0.104*(-0.165,-0.067);0.870**
		K=10	**-1.163*(-1.479,-0.792);0.783**	-0.000(-0.031,-0.000);0.955	-0.115(-1.051,0.219);0.668
		K=15	**-2.022*(-2.481,-1.698);0.766**	-0.000(-0.035,-0.000);0.969	-0.121(-3.337,2.840);0.405
20%	Kmeans	K=5	-0.737(-2.980,0.356);0.091	-0.000(-0.003,-0.000);0.396	-0.000(-0.228,-0.000);0.995
		K=10	-0.422(-1.780,0.848);0.116	-0.000(-0.000,0.002);0.118	-0.000(-1.748,-0.000);0.602
		K=15	-0.168(-1.107,0.235);0.113	0.000(-0.000,0.001);0.090	-0.000(-0.624,-0.000);0.716
	SOM	K=5	-0.490(-1.704,0.594);0.190	-0.000(-0.011,-0.000);0.732	-0.000(-0.188,-0.000);0.999
		K=10	**-0.597*(-1.200,-1.106);0.675**	-0.000(-0.006,-0.000);0.814	-0.000(-0.269,-0.000);0.976
		K=15	**-0.721*(-1.599,-0.078);0.370**	-0.000(-0.011,-0.000);0.731	-0.000(-0.239,0.402);0.640

4.0 CONCLUSION AND DISCUSSION

4.1 CONCLUSION

Based on results, in 3 different down-stream analyses, LRMSE is much well correlated with biological impact measures. Especially, the consistency measure is much higher in DE method, indicating Youden's Index is well-predicted by LRMSE, whereas the impact of MV imputation in classification is even less than DE analysis and the impact of MV treatments in gene clustering depends on a given k value. Thus, in classification and gene clustering analysis, LRMSE is less correlated with biological impact measure when compared to DE method. Thus, superior performance in terms of RMSE based measure does not always guarantee a superior performance in terms of Youden's Index and Adjusted Rand Index. Therefore, the effect on significance analysis, disease classification, and gene clustering such as false discovery rate, the classification accuracy, and clustering consistency need to be taken into account when comparing of ranking of various imputation methods and developing novel imputation methods with RMSE based measure. For now, we emphasize when we select a best imputation method or order ranking of imputation methods, the quality of MV estimates by various imputation methods should be investigated in view of true biological impact as well as RMSE-based measure as many previous studies for MV method have failured to notice this point so far. Additionally, when we estimated MV based on unlogged data prior to log transformation, since the unlogged data set often has a few of large values suspected as outliers, some methods such as LLS-impute, LSA-impute, OLS-impute, PLS-impute, and SVD-Impute sensitively responded to ID(Imputed Data) with some of negative and positive large estimated values, whereas Knn.e-

Impute and KNN.c-Impute, where the imputed values are adjusted by the overall mean and std deviation of the gene and robust imputation method, BPCA are relatively less unstable.(not shown). Thus, if there are some large values on raw dataset, then imputing values could produce outliers as well. This evidence is a result to confirm that log transformation should be performed prior to down-stream analyses.

4.2 FUTURE RESEARCH

To test the relationship between biological impact measure and RMSE-based measure more sophisticatedly, for the consistency measures, some formal hypothesis testing of the hypothesis in 3 major Aims can be done using an ANOVA model based on the simulation results. For the slope estimate, as an alternative to reporting the median and corresponding 95% confidence interval of the simulations, we can fit a model simultaneously to all 100 simulations by using a random effect for the simulation. The model would then include a fixed-effect indicating population mean estimate for the intercept and slope, + random deviations about these due to the simulation random effect, where standard software gives 95% CIs for the fixed effects indicating slope.

BIBLIOGRAPHY

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.G., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L.M., Marti, G.E., Moore, T., Hudson, J., Lu, L.S., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503-511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6745-6750.

Arbeitman, M.N., Furlong, E.E.M., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W. and White, K.P. (2002) Gene expression during the life cycle of Drosophila melanogaster, *Science*, **297**, 2270-2275.

Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M., Misek, D.E., Lin, L., Chen, G., Gharib, T.G., Thomas, D.G., Lizyness, M.L., Kuick, R., Hayasaka, S., Taylor, J.M., Iannettoni, M.D., Orringer, M.B. and Hanash, S. (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, *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289-300.

Bo, T.H., Dysvik, J. and Jonassen, I. (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods, *Nucleic Acids Research*, **32**, 1-8.

BPCA. Available from http://hawaii.aist-nara.ac.jp/~shige-o/tools/

Brock, G.N., Shaffer, J.R., Blakesley, R.E., Lotz, M.J. and Tseng, G.C. (2008) Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes, *Bmc Bioinformatics*, **9**, 1-12.

Burges, C.J.C. (1998) A tutorial on Support Vector Machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, 121-167.

Causton, H.C., Ren, B., Koh, S.S., Harbison, C.T., Kanin, E., Jennings, E.G., Lee, T.I., True, H.L., Lander, E.S. and Young, R.A. (2001) Remodeling of yeast genome expression in response to environmental changes, *Molecular Biology of the Cell*, **12**, 323-337.

de Brevern, A.G., Hazout, S. and Malpertuy, A. (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering, *Bmc Bioinformatics*, **5**,1 -12.

Fraley, C. and Raftery, A.E. (2002) Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611-631.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S.

(1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Hubert, L., Arabie, P.(1985) Comparing Partitions, *Journal of the Classification* 2; 193-218.

Jornsten, R., Wang, H.Y., Welsh, W.J. and Ouyang, M. (2005) DNA microarray data imputation and significance analysis of differential expression, *Bioinformatics*, **21**, 4155-4161.

Kim, H., Golub, G.H. and Park, H. (2006) Missing value estimation for DNA microarray gene expression data: local least squares imputation (vol 21, pg 187, 2005), *Bioinformatics*, **22**, 1410-1411.

Kohonen, T. (2001) Self-organizing maps of massive databases, *Engineering Intelligent Systems for Electrical Engineering and Communications*, **9**, 179-185.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A.M., Tibshirani, R., Botstein, D., Brown, P.O., Brooks, J.D. and Pollack, J.R. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 811-816.

LLS-Impute. Available from http://www.cs.umn.edu/~hskim/tools.html.

LS-Impute. Available from http://www.ii.uib.no/~trondb/imputation/.

Luo, J., Duggan, D.J., Chen, Y.D., Sauvageot, J., Ewing, C.M., Bittner, M.L., Trent, J.M. and Isaacs, W.B. (2001) Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling, *Cancer Research*, **61**, 4683-4688.

Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S. (2003) A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, **19**, 2088-2096.

Ouyang, M., Welsh, W.J. and Georgopoulos, P. (2004) Gaussian mixture clustering and imputation of microarray data, *Bioinformatics*, **20**, 917-923.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C.E., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. (2000) Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, **24**, 227-235.

Scheel, I., Aldrin, M., Glad, I.K., Sorum, R., Lyng, H. and Frigessi, A. (2005) The influence of missing value imputation on detection of differentially expressed genes from microarray data, *Bioinformatics*, **21**, 4272-4279.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203-209.

Simon, R.M., Korn, E.L., McShane, L.M., Radmacher, M.D., Wright, G.W., Zhao, Y.(2004) :Design and Analysis of DNA microarray investigations.

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., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Molecular Biology of the Cell*, **9**, 3273-3297.

The R project for Statistical Computing. Available from htpp://www.R-project.org.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567-6572.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520-525.

Tuikkala, J., Elo, L., Nevalainen, O.S. and Aittokallio, T. (2006) Improving missing value estimation in microarray data with gene ontology, *Bioinformatics*, **22**, 566-572.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001), *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 10515-10515.

van't Veer, L.J., Dai, H.Y., van de Vijver, M.J., He, Y.D.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**, 530-536.

Wang, D., Lv, Y.L., Guo, Z., Li, X., Li, Y.H., Zhu, J., Yang, D., Xu, J.Z., Wang, C.G., Rao, S.Q. and Yang, B.F. (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.

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

Yu, Y.P., Landsittel, D., Jing, L., Nelson, J., Ren, B.G., Liu, L.J., McDonald, C., Thomas, R., Dhir, R., Finkelstein, S., Michalopoulos, G., Becich, M. and Luo, J.H. (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy, *Journal of Clinical Oncology*, **22**, 2790-2799.