A COMPARATIVE STUDY OF DIFFERENT STRATEGIES OF BATCH EFFECT REMOVAL IN MICROARRAY DATA: A CASE STUDY OF THREE DATASETS

by

Fei Ding

B.S. in Astronomy, University of Science and Technology of China, Hefei, China, 2008

M.S. in Physics, University of Pittsburgh, 2010

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Fei Ding

It was defended on

June 7, 2013

and approved by

Thesis Advisor: Yan Lin, Ph.D., Research Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Vincent C. Arena, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Sufi M. Thomas, Ph.D., Assistant Professor, Department of Otolaryngology, School of Medicine, University of Pittsburgh

Yan Lin, Ph.D

A COMPARATIVE STUDY OF DIFFERENT STRATEGIES OF BATCH EFFECT REMOVAL IN MICROARRAY DATA: A CASE STUDY OF THREE DATASETS

Fei Ding, M.S.

University of Pittsburgh, 2013

ABSTRACT

Batch effects refer to the systematic non-biological variability that is introduced by experimental design and sample processing in microarray experiments. It is a common issue in microarray data and could introduce bias into the analysis, if ignored. Many batch effect removal methods have been developed. Previous comparative work has been focused on their effectiveness of batch effects removal and impact on downstream classification analysis. The most common type of analysis for microarray data is differential expression (DE) analysis, yet no study has examined the impact of these methods on downstream DE analysis, which identifies markers that are significantly associated with the outcome of interest. In this project, we investigated the performance of five popular batch effect removal methods, mean-centering, ComBat_p, ComBat_n, SVA, and ratio based methods, on batch effects reduction and their impact on DE analysis using three experimental datasets with different sources of batch effects. We found that the performance of these methods is data-dependent: simple mean-centering method performed reasonably well in all three datasets, but the more complicated algorithms such as ComBat method's performance could be unstable for certain dataset and should be applied with caution. Given a new dataset, we recommend either using the mean-centering method or carefully investigating a few different batch removal methods and choosing the one that is the best for the data, if possible. This study has important public health significance because better handling of batch effect in microarray data can reduce biased results and lead to improved biomarker identification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my sincere gratitude to my thesis advisor, Dr. Yan Lin, for her guidance, encouragement, patience, and input throughout the preparation of this work. I also would like to thank the rest of my committee members, Dr. Vincent Arena and Dr. Sufi Thomas, for their valuable comments and suggestions. Sincere gratitude goes to my parents and friends, whose endless love and support have always been my great companion. A special thanks to my dear husband, Jen-Feng Hsu, for his supporting and encouraging love.

1.0 INTRODUCTION

Microarray techniques have been widely employed in biological and medical research since its invention in the middle 1990s. The ability of processing thousands of probes at one time has brought a revolution to both biological research and statistical analysis of high-throughput data. Many studies require the use of multiple microarrays, with experiments performed at different times, by different technicians, or even at different sites, which introduces batch effects. Here, we refer "batch effects" to any systematic non-biological variability that is introduced by experimental design and sample processing. There are many different sources of batch effects. Some common sources include samples processed at different times, on different chips, at different sites, and by different technicians, and samples coming from differentially processed tissues (e.g. frozen vs. paraffin fixed tissues).

Often, batch effects were ignored in microarray data analysis: Chen et al. pointed out that less than ten percent of 219 papers published in the first half of 2010 addressed batch effects [1]. The goal of employing microarray techniques in biological and medical research is to identify expression heterogeneity among different groups, but the presence of batch effects add variability to expression profile and may lead to biased results. Batch effects still exist even after the microarray signal intensity normalization, so formal removal methods are required to remove batch effects [2].

Many batch effect removal methods have been developed and several papers compared the performances of some of these method [1, 3, 4]. Mean-centering method is a simple ANOVA method that sets the mean of each batch to zero across groups [5]. Standardization method goes one step further beyond mean-centering method: it normalizes the standard deviation within each batch to unity across samples [6]. Ratio-based methods scale expression level by dividing the arithmetic mean (Ratio-A) or geometric mean (Ratio-G) of the control group within each batch. Distance-weighted discrimination (DWD) finds a separating hyper-plane between two batches and projects the batches onto the DWD plane planes, finds the mean, and then subtracts the DWD plane multiplied by this mean [7]. Surrogate variable analysis (SVA) constructs surrogate variables from significant eigenvectors of the residual matrix from which the effect of primary variable has been removed [8]. ComBat (Combating Batch Effects When Combining Batches of Gene Expression Microarray Data) is an empirical Bayes method that includes a parametric prior method (ComBat_p) and a non-parametric method (ComBat_n), and the model includes both additive and multiplicative batch effects [9].

Chen et al. [1] compared six methods, DWD, mean-centering, SVA, Ratio-G, ComBat_p, and ComBat_p, on two simulated datasets and two experimental datasets with batch effects coming from different processing dates and sites. They aimed to assess data integration improvement measured by batch effects reduction, accuracy, precision, and overall performance. Using these four criteria, Chen et al. found that the ComBat performed satisfactorily on all measures, and the mean-centering method was a close second. Other methods had at least one major drawback, for example, the Ratio-G performed worst in removing batch effects from one experimental data. They focused on the “removal” of batch effect regardless of downstream

analysis, which is only one side of the story. The method that removes the batch effect most effectively may “over correct” and mask the true biological signal.

Luo et al. [3] compared five methods, mean-centering, standardization, Ratio-A, Ratio-G, and ComBat, on six datasets from the MAQC-II project [2] with various sources of batch effects including different hybridization time, different generations of chips, different channels, different platforms, and different tissues. Their goal was to evaluate cross-batch prediction performance using the Matthews Correlation coefficient (MCC) as evaluation criteria. They concluded that the ratio-based methods are preferred based on the consensus results of all the 120 cases across six different types of datasets and four different feature selection and classification methods. However, the performance is classifier and data dependent, which is also demonstrated by their results. If different datasets and classification methods were employed, they may reach different results, thus their results may not be generalized to other datasets or different downstream analyses.

As described above, the two published work had focused on different outcomes when comparing a variety of batch effect removal methods. The conclusions of the two papers are inconsistent, indicating that the choice of the methods is dependent on the downstream analysis and different types of datasets. The most common type of analysis for microarray data is differential expression (DE) analysis, where the expression levels of the genes are compared among different groups. Yet, no comparative studies have been done on the impact of batch effect removal methods on DE analysis. In this project, we applied and compared five of the most popular methods, mean-centering, ComBat_p, ComBat_n, SVA, and ratio-based methods, on three array datasets, with batch effects introduced by various sources. We first measured how much each method reduces the variation caused by batch effects using principal variation

component analysis (PVCA). In order to evaluate their impact on DE analysis, we compared the DE analysis results generated from data processed by batch effect removal methods to those generated from original data using other three strategies: (1) Perform DE analysis on original data regardless of batch effects. This is our "negative control". (2) Perform DE analysis on each batch's data (original data without batch effect removal) separately and combine the results using meta-analysis technique. (3) Analyze original data using linear regression and adjust for batch effects as a regression covariate.

2.0 MATERIALS AND METHODS

2.1 DATASETS

Three datasets from three different arrays with different sources of batch effects were used.

2.1.1 Head and neck expression data

This is an Illumina gene expression array data on 29 head and neck cell lines. Profile group is normal cell fibroblast (NNF) vs. tumor cell associated fibroblast (TAF). Two batches were processed at different times: there are 17 (7 TAF and 10 NNF) samples in the first batch, and 12 TAF samples in the second batch.

2.1.2 Melanoma methylation data

Methylation level was assayed by the Illumina H27K array on 65 formalin fixed and paraffin embedded (FFPE) and 19 frozen tumor samples from melanoma patients. Here the batch effects were introduced by different tissue processing methods: frozen and FFPE. This experiment aims to study the methylation profile of tumor tissues with BRAF mutation versus those of wild type. All the 84 samples were used for batch effect removal; however, 24 samples were filtered out for DE analysis (see next section for details).

2.1.3 Lung cancer micro RNA data

It is an Agilent micro RNA (miRNA) expression array data on 120 tumor and 85 normal lung tissue samples of Lung cancer patients. Two batches were processed on two generations of chips, and only the common probes of the two chips were used for analysis. All the 205 samples were used for bath effect removal, however, only the 109 tumor samples with disease-free survival information were used in DE analysis.

2.2 DATA PROCESSING

2.2.1 Normalization and Transformation

The head and neck expression data and lung cancer miRNA data were both quantile normalized. The melanoma methylation was background normalized. Log transformed head and neck expression data was used in all the analysis, and logit transformed melanoma methylation data was used in the regression analysis. The lung cancer miRNA data was analyzed in its original scale.

2.2.2 Missing value imputation

Missing values were imputed by the k-nearest neighbor (KNN) method with k=10 using *impute.knn* function of the R package *impute*.

2.2.3 Filtering

In order to remove markers of poor experimental quality and markers that do not fluctuate across samples, the following filtering procedures were applied to the melanoma methylation data prior to the DE analysis: (1) Only pre-treatment samples were used. (2) Only samples with 75% percentile of detection p-value < 10e-5 were kept. (3) Only sites with median detection p-value < 0.05 were kept. (4) Sites with all average beta values less than 0.2, or greater than 0.8 i.e., maximum average beta value less than 0.2 or minimum greater than 0.8 were removed.

For the lung cancer miRNA data, samples with excessive zero values were removed for DE analysis.

2.3 BATCH EFFECT REMOVAL METHODS

The following batch effect removal methods were considered in this comparative study. Although certain filtering criteria were applied to the data for the DE analysis, we used all available samples and markers for batch effect removal.

2.3.1 Mean-centering

Within each batch, mean expression is calculated across all samples for each gene, and expression level of each sample is adjusted by subtracting the mean expression so that all batches have zero means. It is implemented in the R package *pamr*.

2.3.2 Ratio-based

Within each batch, the geometric mean expression of the reference group is calculated for each gene, and the expression level of each sample is scaled by dividing this geometric mean (referred to as Ratio-G). Arithmetic mean can also be used (referred to as Ratio-A). Luo et al. has shown that Ratio-A is inferior to Ratio-G [3], thus we choose to use Ratio-G in our analysis when possible. The geometric means were calculated by *geometric.mean* function in the R package *Psych*.

2.3.3 SVA

This method combines the method of singular value decomposition (SVD) and the linear model. SVD is applied on a residual expression matrix obtained by removing the effect of the primary variable (here profile group variable) to identify eigengenes, and "surrogate variables" are constructed based on these eigengenes [8]. The main goal of the SVA algorithm is to identify and estimate these surrogate variables.

Let x_{ij} denote expression level of gene *i* on sample *j*, and μ_i denote the baseline level of expression of gene *i*. Let y_j represent the primary variable of interest (e.g. case-control status), and $f_i(y_j)$ gives the relationship between primary variable and expression level of gene *i*. Vectors $\boldsymbol{g_l}=(g_{l1}, \ldots g_{ln})$ are L unmodeled biological and experimental factors. The linear model including primary variable and other unmodeled factors can be written as following,

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{l=1}^{L} \gamma_{li} g_{lj} + e_{ij}^* \quad i = 1, \ldots, m \quad j = 1, \ldots, n \quad l = 1, \ldots, L$$

Instead of directly estimating $\boldsymbol{g_l}$, which is often impossible, an orthogonal set of vectors $\boldsymbol{h_k}$, *k=1,….K* that spans the same linear space as the $\boldsymbol{g_l}$ are identified and the model now can be written as

$$x_{ij} = \mu_i + f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} h_{kj} + e_{ij}^{*} \;\; i = 1, \ldots, m \;\; j = 1, \ldots, n \;\; k = 1, \ldots, K$$

Vectors $\boldsymbol{h_k}$ are called "surrogate variables", and they should be included as covariates in subsequent analyses.

The algorithm can be implemented in five steps: (1) fit a linear model with only primary variable $x_{ij} = \mu_i + f_i(y_j) + e_{ij}$ and calculate the residual expression matrix **R**. (2) Apply SVD to **R** to obtain eigenvalues and eigengenes $\boldsymbol{e}_k$. Evaluate the significance of each eigengene based on their corresponding eigenvalues by a permutation procedure. (3) For each significant eigengene $\boldsymbol{e}_k$, perform a significance analysis of associations to find a subset of $\boldsymbol{m_1}$ genes whose expression levels are associated with this particular eigengene. (4) Apply SVD on the reduced residual matrix and use the eigengene $\boldsymbol{e}_{j*}^{r}$ that is most correlated with $\boldsymbol{e}_k$ to construct a surrogate variable. (5) Include surrogate variables in subsequent analyses.

SVA is implemented in the R package *sva*.

2.3.4 ComBat

ComBat is an empirical Bayes (EB) method developed for adjusting for batch effects in small size data [9]. The model includes both additive and multiplicative batch effects. Let Y_{ijg} represents the expression level of gene g on sample j from batch i. Y_{ijg} can be modeled by a location and scale model as following,

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Where α_g is the overall gene expression, *X* is a design matrix for sample conditions, and β_g is the vector of regression coefficients corresponding to *X*. The error term ε_{ijg} is assumed to follow a normal distribution with mean zero and variance σ_g^2. The γ_{ig} and δ_{ig} represent the additive and multiplicative effect of batch *i* for gene *g*. The batch-adjusted data, Y_{ijg}^* is given by

$$Y_{ijg}^* = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}}{\hat{\delta}_{ig}} + \hat{\alpha}_g + X\hat{\beta}_g$$

Where $\hat{\alpha}_g, \hat{\beta}_g, \hat{\gamma}_{ig}$, and $\hat{\delta}_{ig}$ are estimators of corresponding parameters.

ComBat algorithm is implemented in threes steps.

(1) *Standardize the data* Estimators $\hat{\alpha}_g, \hat{\beta}_g, \hat{\gamma}_{ig}$ were obatained using a gene-wise ordinary least-squares approach, constraining $\sum_i n_i \hat{\gamma}_{ig} = 0$ for all *g=1, …, G*. Estimator of the variance σ_g^2 is given by $\hat{\sigma}_g^2 = \frac{1}{N}\sum_{ij}\left(Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g - \hat{\gamma}_{ig}\right)^2$. The standardized data is calculated by $Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g}$, and it is easy to show that it follows a normal distribution with mean $\gamma_{ig}^{new} = \gamma_{ig}/\sigma_g$ and variance δ_{ig}^2.

(2) *Estimate Parameters* For parametric prior method, assume the batch parameters have the following prior distributions $\gamma_{ig}^{new} \sim N(\gamma_i, \tau_i^2)$ and $\delta_{ig}^2 \sim Inverse\ Gamma\ (\lambda_i, \theta_i)$. Parameters $\gamma_i, \tau_i^2, \lambda_i, \theta_i$ are estimated empirically from standardized data Z_{ijg} using the method of moments. Then the EB estimates for γ_{ig}^{new} and δ_{ig}^2 are given by conditional posterior means. For non-parametric prior method, EB estimates for γ_{ig}^{new} and δ_{ig}^2 are given by estimates of the posterior expectations of the parameters. For both methods, the EB estimates for batch effect parameters are denoted as $\widehat{\gamma_{ig}^*}$ and $\widehat{\delta_{ig}^{2*}}$

(3) *Adjust the data* Using EB estimated batch effects, adjusted data is calculated as

$$Y_{ijg}^{*} = \frac{\hat{\sigma}_g(Z_{ijg} - \hat{\gamma}_{ig}^{*})}{\hat{\delta}_{ig}^{*}} + \hat{\alpha}_g + X\hat{\beta}_g$$

ComBat method is implemented using *ComBat.R* script that can be downloaded at http://www.bu.edu/jlab/wp-assets/ComBat/Download.html

2.3.5 Regression Method

Adjust the batch effects by including the batch as a covariate in the regression model.

2.4 EVALUATION OF BATCH EFFECT REMOVAL METHODS

2.4.1 Effectiveness of batch effect removal

Principle component analysis (PCA) plots were used to visualize batch effect before and after batch effect removal. Principle variance component analysis (PVCA) was employed to measure the amount of variability attributable to batch effect [4, 10]. It combines the methods of principle component analysis (PCA) and variance component analysis (VCA). First, top principle components that explain a proportion of variation just above a preset threshold (60% here) were selected. Then for each retained principle component, fit a mixed model with all factors of interest including interactions as random effects. For each factor, estimate variance components for each model and average estimates across all retained principle components using the corresponding eigenvalues as weights. Finally, weighted average variance components estimates for each factor, interaction term, and the residual variance, were standardized by diving

their sum, thus can be represented as a proportion of the total variance. These proportions can be displayed as bar charts.

2.4.2 Impact on downstream DE analysis

To further evaluate the impact of batch effect removal methods on DE analysis, we compared the results of DE analysis generated by four different strategies: (1) Perform DE analysis on original data regardless of batch effects. (2) Perform DE analysis on each batch's data (original data without batch effect removal) separately and combine the results using meta-analysis technique. (3) Analyze original data using linear regression and adjust for batch effects as a regression covariate. (4) Perform DE analysis on data processed by batch effect removal methods. Top-100 differentially expressed markers were obtained according to the rank of p-values of statistical tests for each strategy. Lacking the knowledge of true DE markers, we used the top list generated by the meta-analysis strategy as our "gold standard". Top-100 markers obtained using other three strategies were compared to the "true" top list generated by the meta-analysis, and a large percentage of overlap with the "true" top list is desirable.

2.4.3 DE analysis

For binary outcomes, e.g., BRAF mutation and wild type, two-sample Wilcoxon Rank Sum test was employed to identify DE markers. For disease-free survival (DFS), the Cox regression was used. The likelihood ratio tests (LRTs) were used to test the association between each marker and the DFS outcome.

2.4.4 Meta-analysis

Fisher's method [11] was employed to combine p-values from individual batch. Specifically, for two-batch case, test statistic $\chi^2 = -2[\log(p_1) + \log(p_2)]$, where $p_1$and p_2 are the p-values from first and second batch respectively, χ^2 has a chi-squared distribution with 4 degrees of freedom, and the p-value can be determined.

3.0 RESULTS

3.1 BATCH EFFECT EVALUATION

PCA was applied to the three datasets before any batch effect removal to visualize batch effect (Figure 1a, 1c, 1e). Using top three principle components (PCs), we can separate different batches perfectly on the PCA score plots for all three datasets.

The PVCA (Figures 1b, 1d, 1f) reveals that before any batch effect removal, main batch effect is very large and it accounts for 50.3%, 39.2%, and 26.8% of the overall variation in three datasets, while main group effect is very small and it only accounts for 2.2%, 2.1%, and 9.3% of the overall variation, respectively. Variability attributable to the interaction term is small in all three dataset and explains 1.1%, 0.9%, and 1.7% of the overall variation.

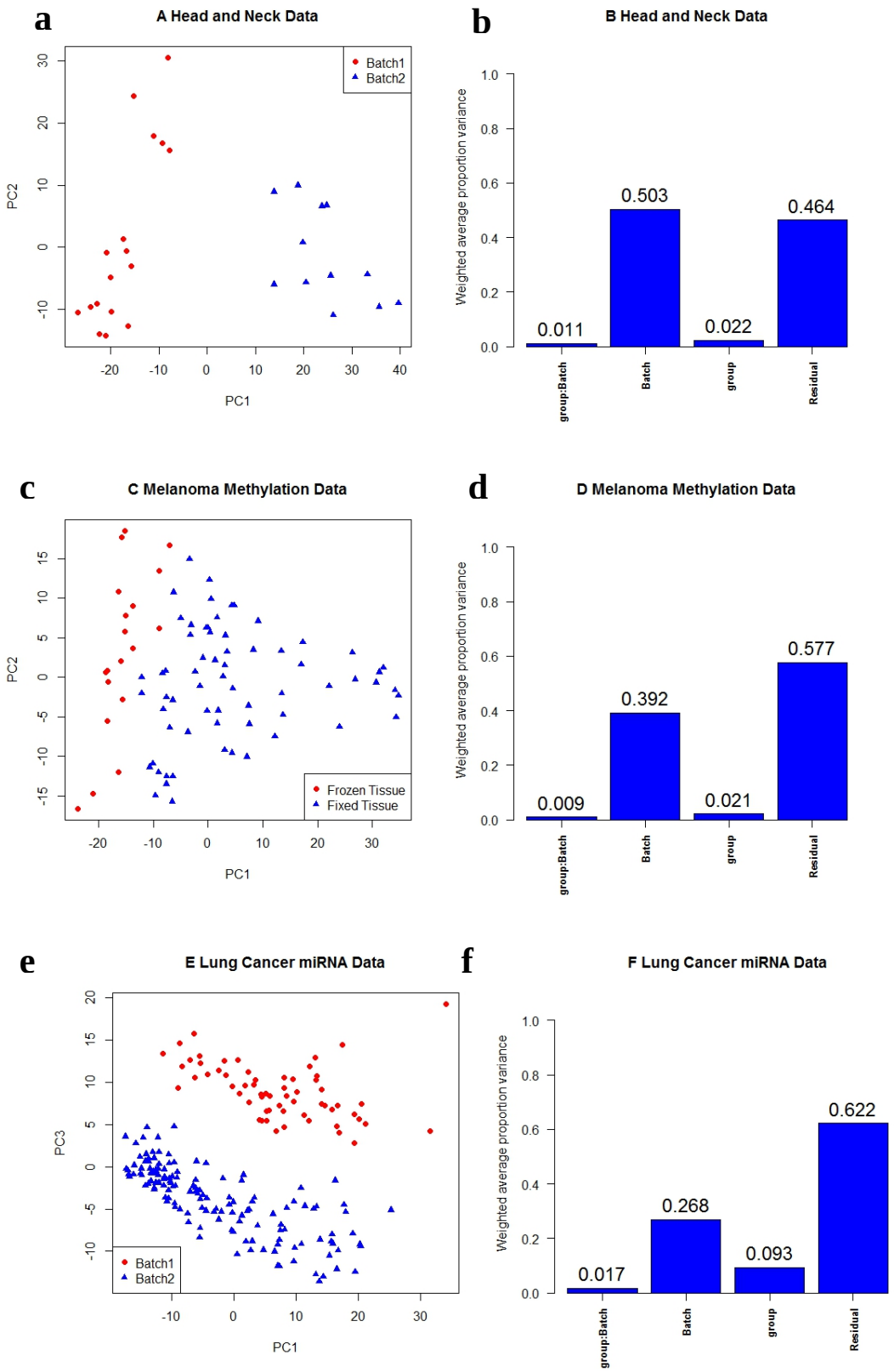

Figure 1. PCA and PVCA results of three datasets before batch effect removal

3.2 BATCH EFFECT REMOVAL EVALUATION

After batch effect removal, batch effects are no longer obvious (see Figure A-4, A-5, and A-6), indicating that all removal methods are capable of removing batch effects to some degree.

To quantitatively measure how much each method reduces batch effects, the PVCA was applied to each dataset before and after batch effect removal (see Figure A-7, A-8, and A-9 for details). As described in materials and methods section, PVCA is a method combining principle component analysis and variance component analysis to estimate how much variation in the expression data is attributable to batch effects and other factors. Here we consider three factors: main batch effect, main profile group effect, and the interaction between batch and group effects. Group effect is the factor of interest, and we want to study differences in expression profile induced by it; while batch effect is the non-biological differences when samples are processed in different batches, and we want to get rid of it. The PVCA results were summarized in Figure 2 to demonstrate the variation due to (a) main batch effect, (b) batch related factors, i.e., main batch effect and the interaction term considered together, (c) main group effect, (d) group related factors, i.e., main group effect and the interaction term considered together before and after the batch effect removal for each dataset.

The PVCA revealed that main batch effect and the interaction term together explained 51.4% of the overall variation in the head and neck expression data without batch effect removal. All four batch removal methods reduced that variation to less than 2%, and ComBat_n and ComBat_p eliminated it completely. This reduction made the biological variation due to main group effect more apparent, increasing it from 2.2% to 3.2% (mean-centering), 15.3% (ComBat_p), 14.6% (ComBat_n), and 4.4% (SVA) of the overall variation after batch effect removal (Figure A-7). Notice that mean-centering method increased the variation due to the

interaction term from 1.1% to 1.6%, but this increase was not as serious as the ones we observed later. Ratio-based methods were not applied on this dataset because the second batch did not have a reference group.

Main batch effect and the interaction term explained 39.2% and 0.9% of the overall variation in the melanoma methylation data, respectively before batch effect removal. Only the mean-centering method was able to reduce both of these two variations to 0.8% and 0.4%, respectively. All the other methods either greatly increased the variation attributable to the interaction term, e.g., ComBat_n increased it to 12.4%, or was not very effective in reducing main batch effect, e.g. main batch effect still accounted for 15.4% of the overall variation for the data processed by SVA (Figure A-8). The signal of this dataset is relatively weak. After applying mean-centering method to the data, the variation attributable to main group effect increases from 2% to 4%. Although this variation was also increased to 4.1% by the ComBat_n method, it was still far less than 12.4%, the variation due to the interaction term. For ComBat_p, SVA, and ratio-G, increase in variation due to group related factors were mostly because of the increase in the interaction terms instead of the increase in main group effect.

For the lung cancer miRNA data, the ComBat_p, and the ComBat_n methods increased variation attributable to the interaction term even more severely. This variation accounted for over 35% of the total variation in ComBat_p processed data and over 25% of the total variation in ComBat_n processed data (Figure A-9). For this dataset, the mean-centering and the SVA methods performed better: they produced data with reduced variations attributable to both main batch effect and the interaction term and increased variation due to the main group effect, with the latter doing better. The Ratio-A method didn't reduce batch effects as effectively as the

previous two methods. We used the Ratio-A instead of the Ratio-G method for this dataset because there were samples with zero values.

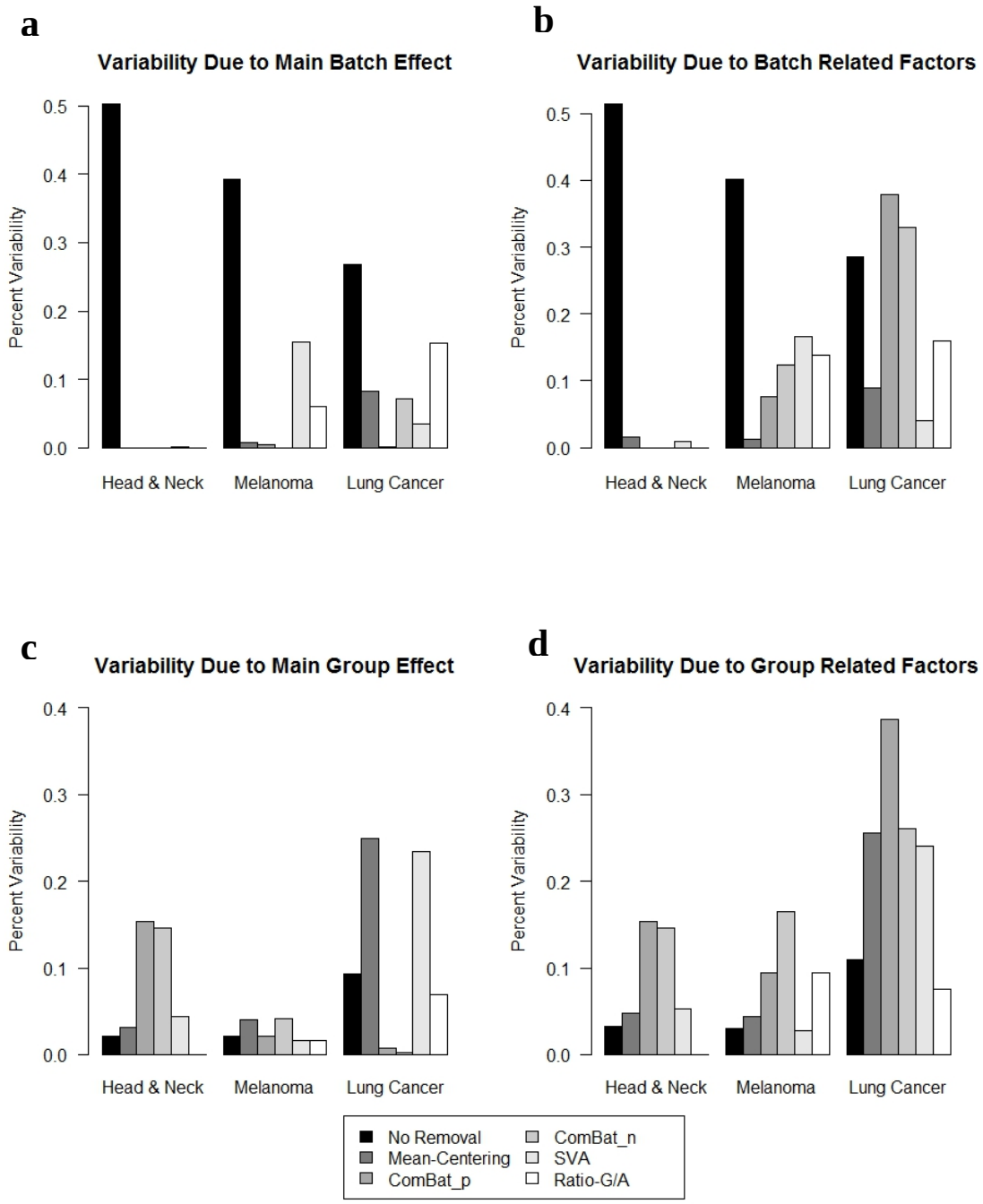

Figure 2. PVCA results of three datasets before and after batch effect removal

3.3 DE ANALYSIS RESULTS

To study the impact of different batch effect removal methods on the DE analysis, we compare the results of the following four strategies. (1) Perform DE analysis on original data regardless of batch effects. This is our "negative control". (2) Perform DE analysis on each batch (original data without batch effect removal) separately and combine the results using meta-analysis technique. This is our "gold standard". (3) Analyze original data using linear regression and adjust for batch effects as a regression covariate. (4) Perform DE analysis on data processed by batch effect removal methods. Here, we focused on one simple batch effect removal method, mean-centering, and one complex method, ComBat_n. SVA was also evaluated for the lung cancer miRNA data for its good performance on that data.

Top-100 differentially expressed marker list was obtained according to the rank of p-values for each strategy. For comparison purpose, results of the individual batch data are also included. We refer the top list produced by meta-analysis as the "true" top list and recorded the percent of markers on the top-100 list generated by other strategies that are also on the true top marker list. We consider a larger percentage of overlap between the top lists generated by other strategies and this true top marker list an indication of more reliable results.

Given the experimental design used in the study that provided the head and neck expression data, it was not included for DE analysis.

3.3.1 Analysis of the melanoma methylation dataset

As stated in materials and methods section, the melanoma methylation data was filtered prior to the DE analysis, and there are 18303 candidate sites and 60 samples left. The results were

summarized in Table 1. Without any batch effect removal/control, 38% of the top markers generated were also on the true top marker list. When we analyzed the frozen tissue batch and the FFPE tissue batch separately, we observed a 21% and 43% overlap with the meta-analysis results, respectively. This is expected given the small number of frozen tissues (19 samples) and relatively large number of FFPE tissues (41 samples). We also noticed that none of the top-100 marker lists of these two separate analysis overlapped, demonstrating the differences between these two types of tissues. The overlap between the top marker lists generated by different batch effect removal/control methods and the true top lists ranged from 34-49%, with the simple mean-centering processed data giving the most reliable list. Overall this dataset has very weak signal: almost all the strategies gave zero significant markers controlling false discovery rate (FDR) at 20% (q-value<0.2). One interesting thing we noticed is that for the ComBat_n processed data, we obtained 128 significant associated markers controlling FDR at 20%. However, given the fact that other than residual variation, most of the variation is attributable to the interaction term (see Figure A-8d), we cannot distinguish the batch effect and the group effect well. For this reason, we think these discoveries are likely to be false discoveries. Overall, the mean-centering method performed most satisfactorily for this dataset.

Table 1. Comparison of DE results for the melanoma methylation dataset

Method	Percentage of overlap with "true" top-100 marker list	Number of markers with q-values<0.2
Meta-Analysis (60 samples)	Gold Standard	0
Frozen Batch Only (19 samples)	21	0
FFPE Batch Only (41 samples)	43	0
Original Data (60 samples)	38	0
Mean-centering Processed Data (60 samples)	49	0
ComBat_n Processed Data (60 samples)	34	128
Regression (60 samples)	42	0

3.3.2 Analysis of the lung cancer Micro RNA dataset

For the lung cancer miRNA data (Table 2), pairs of normal and tumor tissues were sampled. We used the normal vs. control status as the profile group effect when running batch effect removal algorithms and the subsequent PVCA analysis. This is reasonable because this variable probably affects the miRNA expression level more than any other factors that are measured in this dataset.

However (and fortunately), paired samples of the same patient were always assayed in the same batch in our data. Thus for the comparison between the tumor and normal tissue, batch effect would be automatically taken care of by the subtraction of expression levels of paired samples coming from the same person. Therefore, for our purpose of comparing different batch removal methods, we chose to use the 109 tumor samples at baseline and conducted DE analysis to look for markers associated with disease free survival (DFS) outcome. The cox regression and likelihood ratio tests were used. Again, using meta-analysis result as our gold standard, results of the batch1 data (46 samples) overlapped with 59% of true top list, and results of the batch 2 data (63 samples) overlapped with 53% of true top list. When data from the two batches were combined without any batch removal, resulted top differentially expressed marker list overlapped with only 30% of the true top list. None of the batch effect removal/control methods were able to improve this result significantly: we found 32% for the mean-centering processed data, 34% for both ComBat_p and ComBat_n processed data, 32% for the SVA processed data, and 39% for the regression method. This result is somewhat surprising. It indicates that the effects of each marker on DFS in the two batches are qualitatively different, i.e. the directions of the effect size of some of the markers on the top lists are different in these two batches. Among the 100 markers on batch 1 only analysis top list, 64 showed an association with the DFS of opposite direction in batch 2 data. Similarly, among the 100 markers on batch 2 only analysis top list, 73 showed an association with the DFS of different direction in batch 1 data. In addition, the time-to-event endpoint is often tricky to deal with at the presence of batch effects. Luo et al. also noted that batch removal methods may not improve performance if time to event endpoint was used [3].

Table 2. Comparison of DE results for the lung cancer micro RNA dataset

Method	Percentage of overlap with "true" top-100 marker list	Number of markers with q-values<0.2
Meta-Analysis (109 samples)	Gold Standard	0
Batch1 Only (46 samples)	59	0
Batch2 Only (63 samples)	53	0
Original Data (109 samples)	30	0
Mean-Centering Processed Data (109 samples)	32	0
ComBat_p Processed Data (109 samples)	34	0
ComBat_n Processed Data (109 samples)	34	0
SVA Processed Data (109 samples)	32	2
Regression (109 samples)	39	0

4.0 DISCUSSION

Batch effect is a common issue in microarray data, and it could introduce bias into the analysis, if ignored. Previous comparative work has been focused on effectiveness of batch effect removal [1] and impact on downstream classification analysis [3]. In this project, we investigated the performance of five popular batch effect removal methods, mean-centering, ComBat_p, ComBat_n, SVA, and ratio based methods, on batch effects reduction and their impact on DE analysis using three experimental datasets with different sources of batch effects.

We found that the performance of these methods is data dependent. All four methods were able to remove batch effects in the head and neck expression data effectively, with Combat methods slightly outperforming the others. On the other hand, the Combat methods didn't perform well in either of the remaining two datasets in terms of reducing the variation due to the interaction between batch and group effect. For the melanoma methylation data, the mean-centering method performs the best regarding both batch effect reduction and the DE analysis. The SVA removes batch effect most effectively in the lung cancer micro RNA dataset, followed by the mean-centering method. Looking across the three datasets, we noticed that the mean-centering method consistently generates reasonable results while the Combat methods, although more sophisticated, failed on two out of the three datasets. SVA method's performance also heavily depends on the dataset: it was effective in two out of the three datasets, but not the other one. Ratio-G (or Ratio-A) didn't stand out in either of the two datasets that we applied it to,

probably due to the fact that our experiment is not ideally designed for this method. If we have true internal control that was run in every batch, we will expect to see much better performance of this method.

Similar to Chen et. al 2011 [1], we observed a better biological signal after the batch effect removal represented by an increase in the variability due to main group effect in most cases (Figure 2c). However, theoretically, it could go either way depending on the distribution of the cases and the magnitudes and directions of the group and batch effects. For example, when the true signal is very weak, we may end up with group effects that is slightly higher or lower after the batch effect removal (Figure 2c).

To investigate the impact of batch effect removal methods on DE analysis, we compared the top lists generated by different strategies to the "true" top list generated by the meta-analysis of the two batches. Interestingly, for the methylation data, the mean-centering processed data generated the most reliable results followed by simple regression adjustment of batch effect. Of concern, the Combat_n method seemed to increase the power greatly by declaring 128 significant markers (compared to none for all the other methods) at FDR=0.2. However, given the relatively large variation attributable to the interaction between the batch and the group effect, the signals here are most likely to be false positive. This behavior for the Combat method is also observed in the lung cancer micro RNA dataset. Therefore, we should be cautious when using the more complicated model based batch effect removal methods.

Given a new dataset, we recommend either using the mean-centering method or carefully investigating a few different batch removal methods and choosing the one that is the best for the data, if possible.

In our analysis of the melanoma methylation data, we noted an interesting result. We started with a total of 58 FFPE samples. However, 17 FFPE samples didn't meet our quality control filtering. When we included these samples in the analysis, a large proportion (>70%) of the markers became significant (FDR=0.2). When we take a close look at the PCA plots of the FFPE samples, we can clearly see that these 17 samples separate from the rest of the samples (the green and yellow points on Figure 3). More interestingly, these samples also happened to separate the two groups perfectly, which caused the large amount of significant signals. This interesting result emphasized the importance of appropriate preprocessing of data.

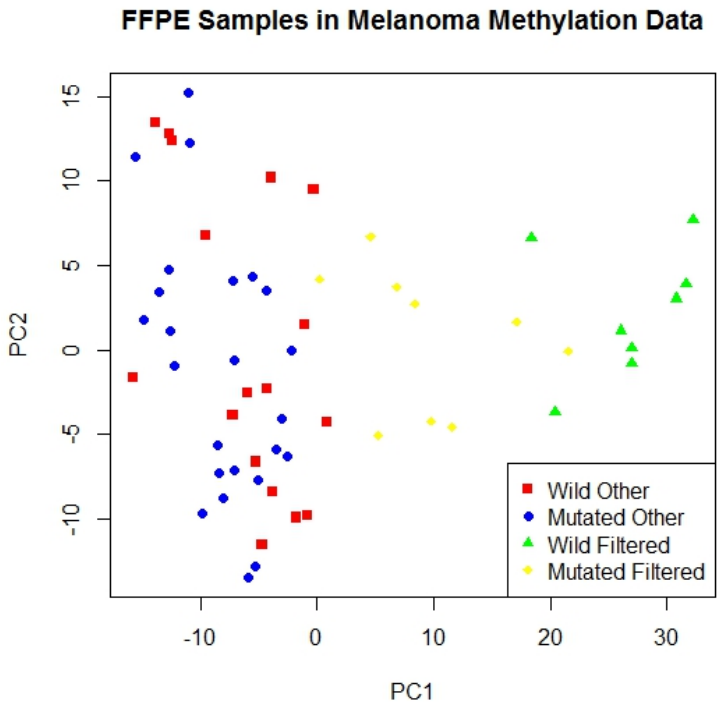

Figure 3. PCA score plot of the FFPE samples in the melanoma methylation dataset

Last, we would like to shed some light on the study design. When the only option is to run the samples in multiple batches, it is important to balance the cases and controls across the

batches and make sure that sample are randomly assigned to the batches to avoid any systematic differences between the samples assigned to different batches. When cases and controls are completely separated, we simply cannot distinguish the group effect from the batch effect. For paired study design, e.g. our lung cancer micro RNA dataset, running the paired samples from the same subjects in the same batch will efficiently minimize the batch effects. The inclusion of technical replicas across different batches is also important in the assessment and the correction of batch effects.

4.1 LIMITATION AND FUTURE WORK

The main limitation of this work lies in the lack of "true positives". As a compromise, we used the meta-analysis results as our "gold standard" to evaluate the impact of batch effect removal methods on DE analysis. However, the method we used to combine p-values, the Fisher's method, has its own limitations: it can be dominated by extreme small p-values in one batch; in addition, it doesn't count for different directions of effect sizes. This becomes a problem especially for our lung cancer micro RNA data. In the future we plan to conduct simulation studies. Specifically, we will spike in true signals on null data generated from permutation of real datasets so that we know the true positives.

APPENDIX A

ADDITIONAL FIGURES

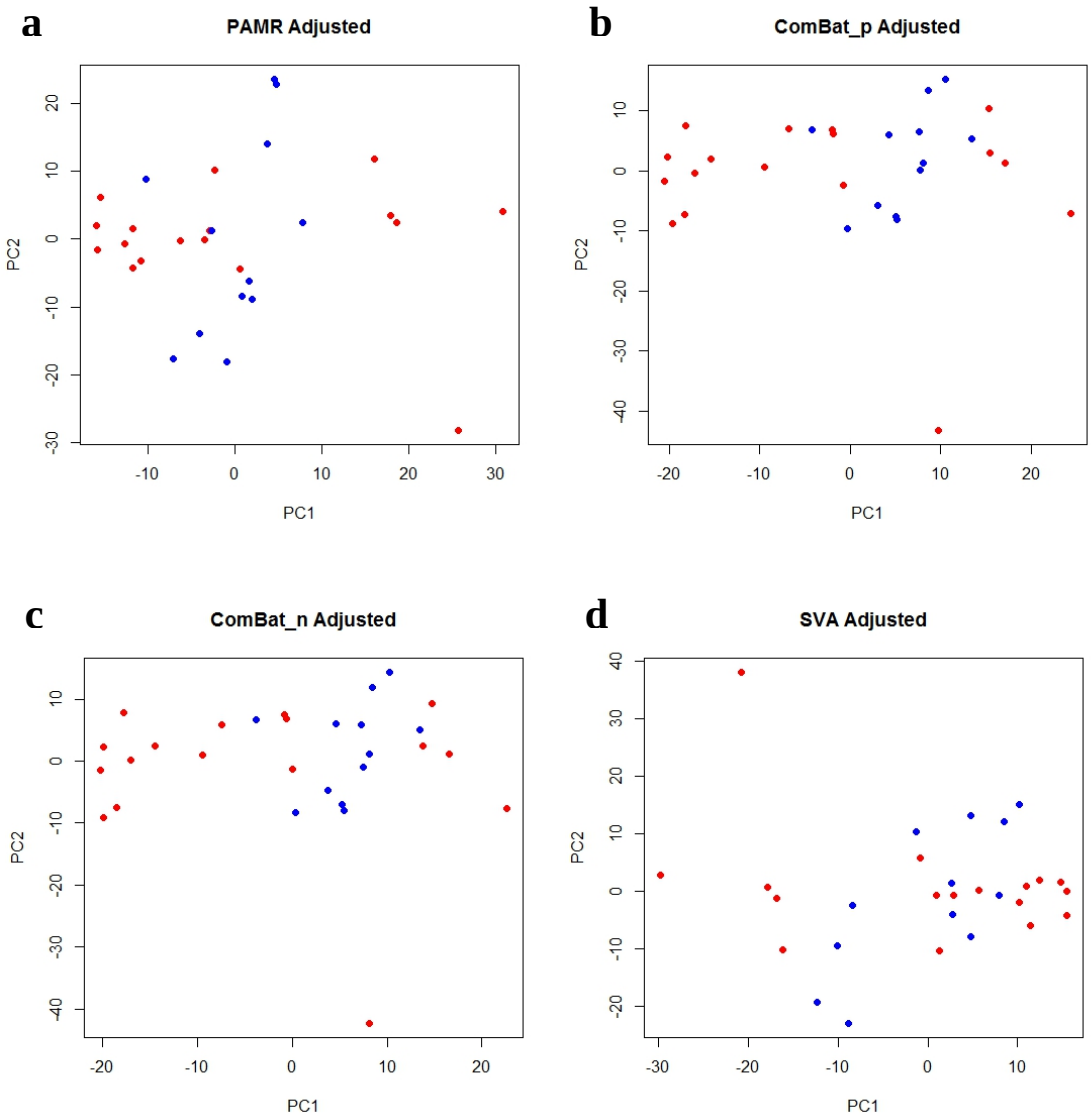

Figure A-4. PCA score plots for head and neck expression data

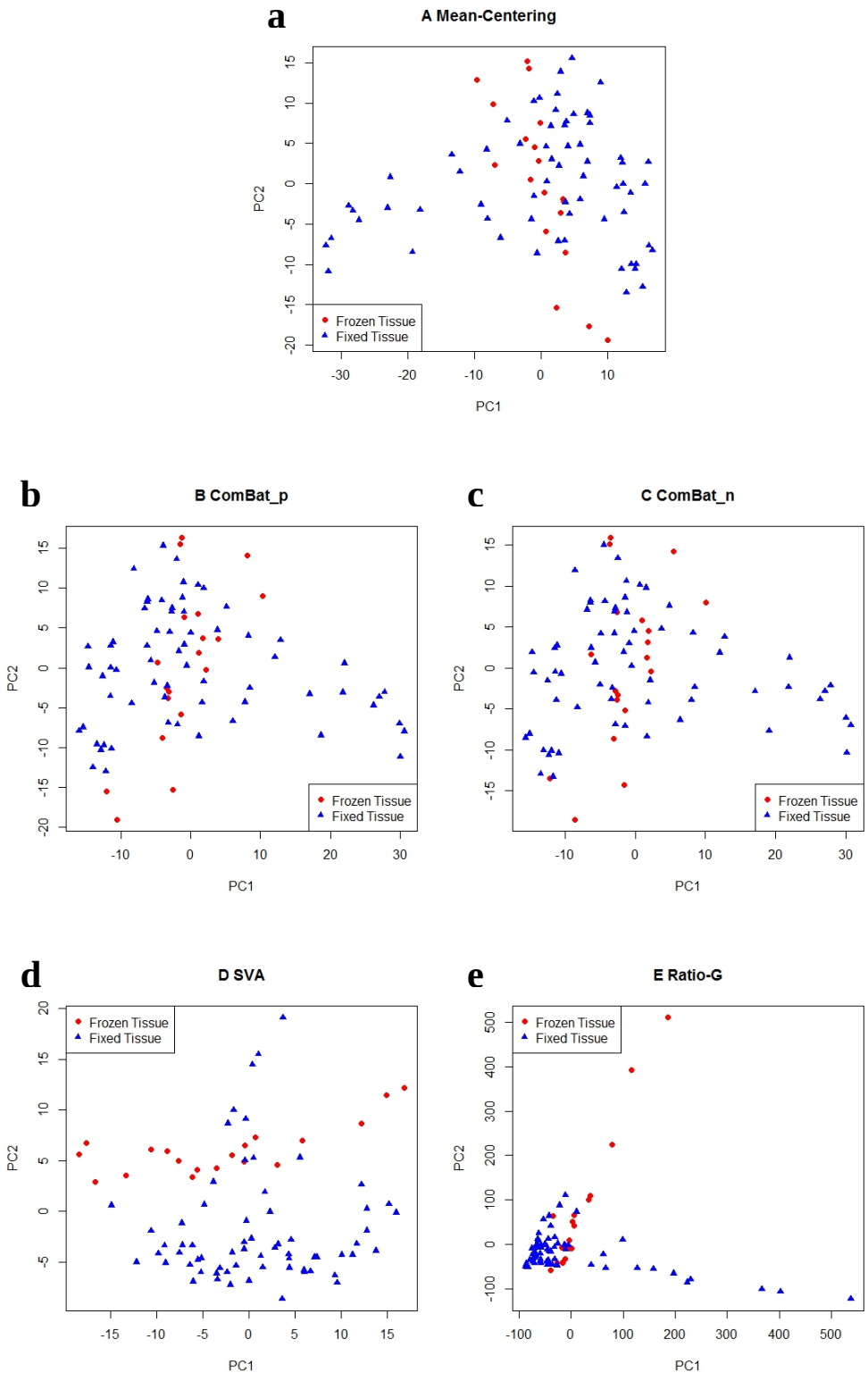

Figure A-5. PCA score plots for melanoma methylation data

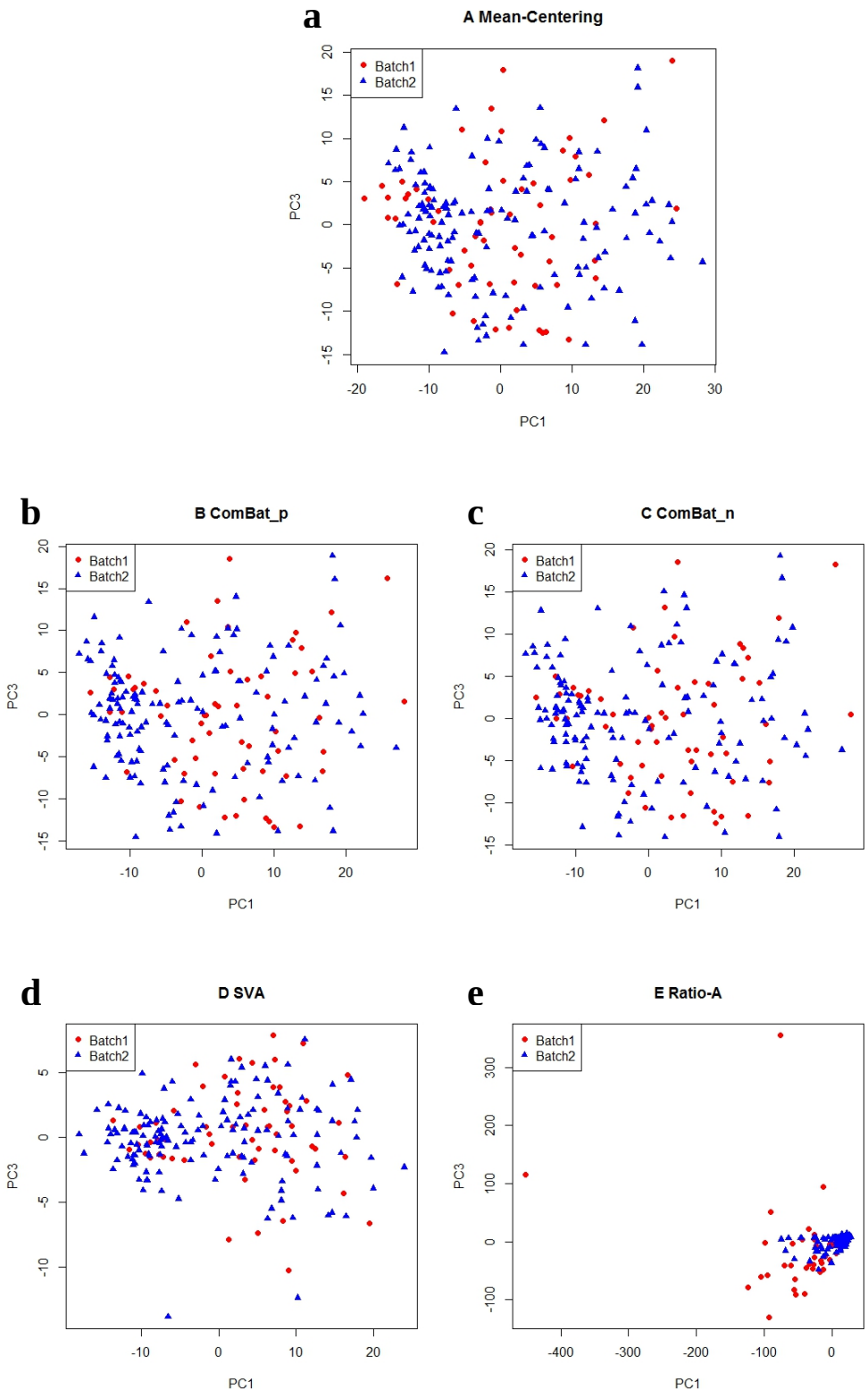

Figure A-6. PCA score plots for lung cancer miRNA data

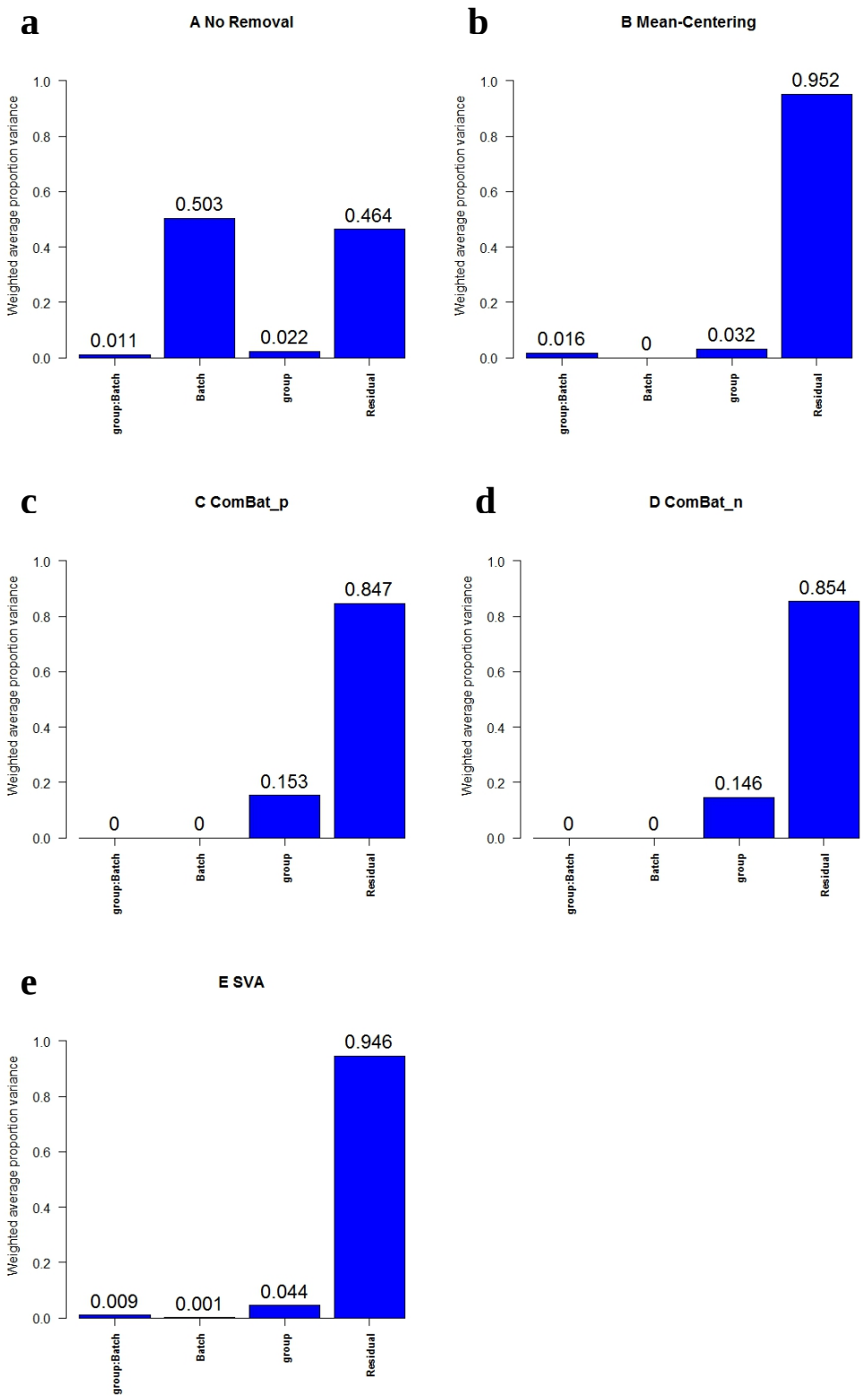

Figure A-7. PVCA results in head and neck expression data

.

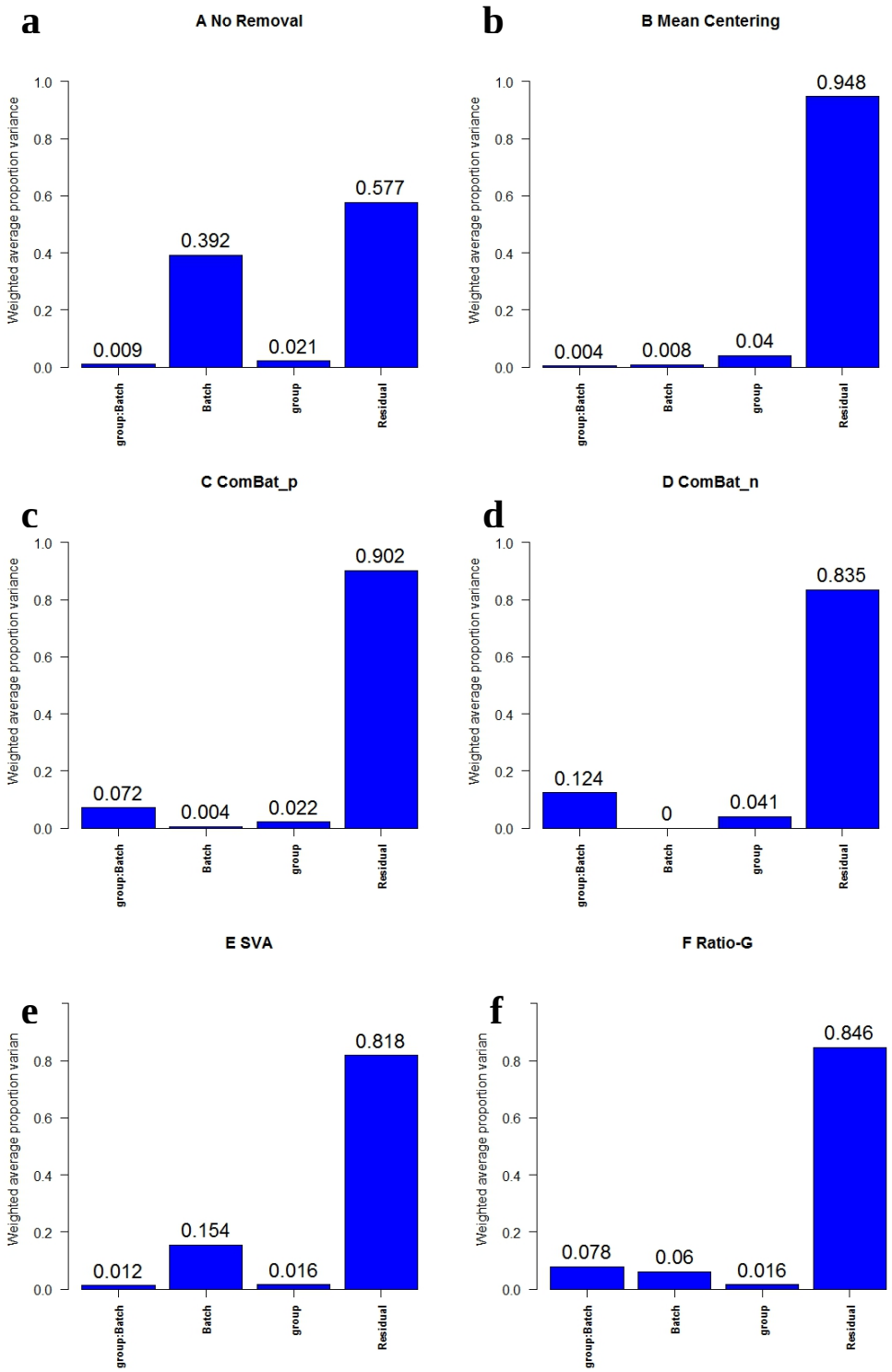

Figure A-8. PVCA results in melanoma methylation data

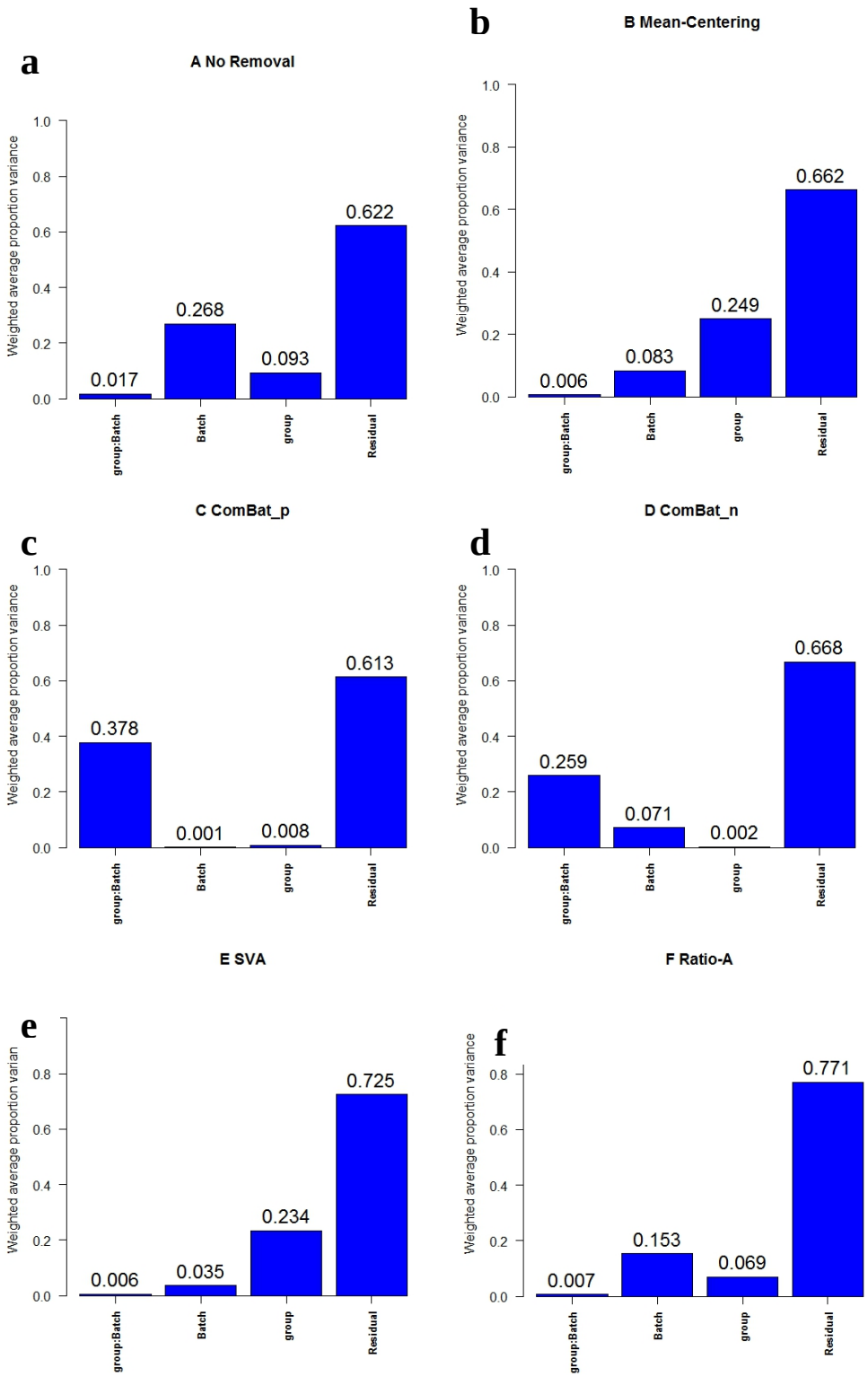

Figure A-9. PVCA results in lung cancer miRNA data

APPENDIX B

ABBREVIATIONS

ComBat_n	non-parametric ComBat method
ComBat_p	parametric ComBat method
DE	differential expression
DFS	disease-free survival
DWD	distance-weighted discrimination
EB	empirical Bayes
FDR	false discovery rate
FFPE	formalin fixed and paraffin embedded
KNN	k-nearest neighbor
LRT	likelihood ratio test
MAQC-II	MicroArray Quality Control Consortium Phase II
MCC	Matthews Correlation Coefficient
NNF	normal cell fibroblast
PC	principal component
PCA	principal component analysis
PVCA	principal variance component analysis
Ratio-A	ratio-based method (using arithmetic mean)
Ratio-G	ratio-based method (using geometric mean)
SVA	surrogate variable analysis
SVD	singular value decomposition
TAF	tumor cell associated fibroblast
VCA	variance component analysis

BIBLIOGRAPHY

1. Chen, C., et al., *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.* PLoS One, 2011. **6**(2): p. e17238.

2. Shi, L., et al., *The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models.* Nat Biotechnol, 2010. **28**(8): p. 827-38.

3. Luo, J., et al., *A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.* Pharmacogenomics J, 2010. **10**(4): p. 278-91.

4. Scherer, A., *Batch effects and noise in microarray experiments : sources and solutions*. Wiley series in probability and statistics. 2009, Chichester, U.K.: J. Wiley. xx, 252 p.

5. Sims, A.H., et al., *The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets - improving meta-analysis and prediction of prognosis.* BMC Med Genomics, 2008. **1**: p. 42.

6. Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.* Proc Natl Acad Sci U S A, 2001. **98**(1): p. 31-6.

7. Benito, M., et al., *Adjustment of systematic microarray data biases.* Bioinformatics, 2004. **20**(1): p. 105-14.

8. Leek, J.T. and J.D. Storey, *Capturing heterogeneity in gene expression studies by surrogate variable analysis.* PLoS Genet, 2007. **3**(9): p. 1724-35.

9. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

10. Boedigheimer, M.J., et al., *Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories.* BMC Genomics, 2008. **9**: p. 285.

11. Fisher, R.A., *Statistical methods for research workers*. 1925, Edinburgh, London,: Oliver and Boyd. ix p., 1 l.