Published online 10 January 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 1 **1**
https://doi.org/10.1093/nargab/lqac100

On data normalization and batch-effect correction for tumor subtyping with microRNA data

Yilin Wu[1,†], Becky Wing-Yan Yuen[1,†], Yingying Wei[2] and Li-Xuan Qin[1,*]

[1]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA and [2]Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, SAR, China

Received April 28, 2022; Revised October 21, 2022; Editorial Decision November 01, 2022; Accepted December 12, 2022

ABSTRACT

The discovery of new tumor subtypes has been aided by transcriptomics profiling. However, some new subtypes can be irreproducible due to data artifacts that arise from disparate experimental handling. To deal with these artifacts, methods for data normalization and batch-effect correction have been utilized before performing sample clustering for disease subtyping, despite that these methods were primarily developed for group comparison. It remains to be elucidated whether they are effective for sample clustering. We examined this issue with a re-sampling-based simulation study that leverages a pair of microRNA microarray data sets. Our study showed that (i) normalization generally benefited the discovery of sample clusters and quantile normalization tended to be the best performer, (ii) batch-effect correction was harmful when data artifacts confounded with biological signals, and (iii) their performance can be influenced by the choice of clustering method with the Prediction Around Medoid method based on Pearson correlation being consistently a best performer. Our study provides important insights on the use of data normalization and batch-effect correction in connection with the design of array-to-sample assignment and the choice of clustering method for facilitating accurate and reproducible discovery of tumor subtypes with microRNAs.

INTRODUCTION

Accurate tumor subtypes are needed to facilitate disease diagnosis, prognosis and treatment in clinical oncology (1). Recent decades have witnessed new and improved tumor subtyping afforded by transcriptomics data via cluster analysis (2–5). However, some of the published subtypes were later found to be not reproducible (6–8). The irreproducibility can be partly attributed to the ubiquitous artifacts in transcriptomics data that arise from disparate experimental handling (9–11). While such artifacts are typically managed with between-sample 'normalization' and across-batch 'correction' in data preprocessing, it relies on borrowing normalization and correction methods that were developed and validated for differential expression analysis comparing two sample groups (12–15). As we recently showed, these methods behave differently when the analysis goal changes to sample classification and survival prediction (16–19). To date, very limited research has been done on their behavior when the analysis goal is sample clustering for tumor subtyping (20). Furthermore, it remains to be elucidated how data normalization and batch-effect correction (BEC) impact cluster analysis for microRNAs (miRNAs), an important class of small RNAs that regulate gene expression and are believed to play an important role in tumorigenesis (21–23).

We set out to study the role of data normalization and BEC in sample clustering for miRNA data. Leveraging a pair of miRNA microarray data sets that were previously collected for tumor samples of two histological subtypes, we conducted a simulation study using a re-sampling algorithm that generated realistically distributed miRNA data under various levels of biological signals and handling artifacts (24,25). In this article, we report our findings for three popular normalization methods and one popular BEC method as well as their combinations, when used along with seven clustering methods, to glean important insights for reproducible tumor subtyping.

MATERIALS AND METHODS

Here, we introduce briefly the empirical and simulated data and describe the methods for normalization, BEC and clustering that we evaluated.

Empirical data collection

A set of 192 untreated primary gynecologic tumor samples (96 endometrioid endometrial tumors and 96 serous ovarian tumors) were collected at Memorial Sloan

*To whom correspondence should be addressed. Tel: +1 646 888 8251; Email: qinl@mskcc.org
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Kettering Cancer Center from 2000 to 2012. They were profiled using the Agilent Human miRNA Microarray (Release 16.0, Agilent Technologies, Santa Clara, CA), following the manufacturer's protocol. This array platform contains 3523 markers (representing 1205 human and 142 human viral miRNAs) and multiple replicates for each marker (ranging from 10 to 40). Two datasets were obtained from the same set of tumor samples using different methods of experimental handling. The first dataset (hereafter referred to as the uniformly handled dataset) was handled by one technician in one batch with the arrays assigned to the samples using blocked randomization (as the arrays come in eight-plex array slides serving as 'blocks' of experimental units); its data quality was further validated using qPCR and also technical replicate arrays. By contrast, the second dataset (hereafter referred to as the non-uniformly handled dataset) was collected by two technicians over multiple batches in the order of sample collection; the first 80 arrays were collected by one technician in two batches and the last 112 by a second technician in three batches. More details on data collection can be found in Qin *et al.* (24–26).

Re-sampling-based data simulation

Additional data were numerically generated by (i) estimating biological effects for the tumor samples (that is, the 'virtual samples'), (ii) estimating handling effects for the arrays in the non-uniformly handled dataset (that is, the 'virtual arrays') and (iii) virtually assigning and then hybridizing the virtual arrays and the virtual samples. More specifically, first, we used the uniformly-handled dataset to approximate the 'biological effects' for each tumor sample; second, for each sample we used the difference between its two arrays from the two datasets to approximate the 'handling effects' for its array in the non-uniformly-handled dataset; lastly, data were simulated by assigning virtual arrays to virtual samples under a study design (either partially confounded or stratified) and summing the biological effects for a virtual sample and the handling effects for its assigned virtual array. A partially confounded design assigned 90% of the first 96 arrays and 10% of the last 96 arrays to ovarian samples and the rest of the arrays to endometrial samples. A stratified design assigned arrays in each batch to the two tumor groups in equal proportions. This simulation strategy assumed an additive model on the log scale for biological effects and handling effects: the uniformly-handled design made every effort to minimize handling effects, and its data was considered to be a best approximation of biological signals; on the other hand, the data collected by the non-uniformly-handled design resembling typical practice exhibited excessive handling effects, and its difference from the uniformly-handled data were used to approximate the noises due to disparate handling.

Simulation scenarios

Varied levels of biological signals. Comparing the two tumor subtypes using the estimated biological effects, 351 (10%) out of the total of 3523 markers were significantly differentially expressed ($P < 0.01$). By chance, 35 markers were expected to have a P-value <0.01. We simulated various levels of biological signals by varying the proportion and magnitude of differential expression. The proportion of differential expression (denoted as π) was varied by removing a portion of non-differentially expressed markers; the magnitude of differential expression for significant markers was varied via amplifying their group mean differences by a constant (denoted as c). That is, for a marker whose group means in ovarian and endometrial samples are μ_1 and μ_2, respectively, $c^*|\mu_1 - \mu_2|$ was added to the group with the larger mean.

Varied magnitudes of experimental artifacts. We introduced various levels of data artifacts through amplifying estimated handling effects by a constant (denoted as d). In this study, we first examined eight settings of (π, c) for the biological-effects only data, where π is 10% or 30% and c is 0, 0.8, 1.6 or 2.4. Based on their clustering results, we chose six settings of (π, c, d)—(10%, 0.8, 1), (10%, 1.6, 1), (10%, 2.4, 1), (30%, 0.8, 1), (30%, 1.6, 1) and (10%, 1.6, 3)—for generating data with both biological and handling effects.

Preprocessing of the simulated data

The preprocessing of each simulated dataset followed three steps: (i) median summarization for the replicate probes of each marker; (ii) data normalization with or without BEC and (iii) $\log_2$ transformation (27). For BEC, we used the ComBat method, with the batch variable defined as the array slide for biological-effects only data and as the array slide or the handling batch for handling-effects-added data (28). To our knowledge, ComBat is the most popular BEC method when the batch variable is known, which is the case for our data. For normalization, we examined three methods that are relatively commonly used in the literature (29–31).

- Median normalization. This method is to shift the data of each sample by an additive constant so that their median becomes the same across samples (32).
- Quantile normalization. This method is to equate the rank statistics across samples (33). We carried out the computation by the *normalize.quantiles( )* function from the R package *preprocessCore* (34).
- Variance stabilizing normalization (VSN). This method was proposed by Huber *et al.* (35). The idea is to tackle the dependence between mean and standard deviation of the data by parametric transformations. We carried out the computation by the *vsn2( )* function from the R package *vsn*.

Clustering of the preprocessed simulated data

The preprocessed data were inputted to sample clustering using both algorithm-based methods and model-based methods. The former included *K*-means clustering, Sparse *K*-means and Partition Around Medoids (PAM); the latter included Self-Organizing Map (SOM) and the Multivariate Normal Mixture (MNM) model.

- *K*-means, proposed by Lloyd and Forgey, is a well-known clustering algorithm (36,37). The general idea is to separate samples into K clusters by assigning them to the near-

Table 1A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 2.4) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	1.00	1.00	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	1.00	1.00	1.00	0.98
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	1.00	1.00	1.00	0.96
SOM	0.98	0.98	1.00	1.00
MNM	0.01	0.68	0.96	0.94

Table 1B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 2.4) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	1.00	1.00	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	0.98	1.00	1.00	1.00
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	0.94	0.98	0.98	0.98
SOM	1.00	1.00	1.00	1.00
MNM	0.96	0.96	0.98	0.96

Table 2A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 1.6) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	0.98	0.98	1.00	1.00
Sparse *K*-means	1.00	1.00	0.98	1.00
PAM Euclidean	0.98	1.00	1.00	0.96
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	0.96	0.96	0.96	0.94
SOM	0.98	0.98	0.98	0.98
MNM	0.01	0.00	0.88	0.00

Table 2B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 1.6) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	1.00	0.98	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	0.98	0.98	1.00	0.92
PAM Pearson	1.00	1.00	1.00	0.96
PAM Spearman	0.90	0.92	0.92	0.80
SOM	0.98	0.98	1.00	0.98
MNM	0.00	0.84	0.88	0.76

est centroids in an iterative manner. We carried out the computation by the build-in kmeans() function in R.

- Sparse *K*-means, proposed by Witten and Tibshirani, utilizes a lasso-type penalty to select a subset of features adaptively for clustering (38). We carried out the computation by the KMeansSparseCluster() function from the R package sparcl (39).
- PAM, a modified version of *K*-means proposed by Kaufman and Rousseeuw, assigns samples to their nearest medoids instead of centroids (40). We carried out the computation by the pam() function from the R package cluster and considered three choices of distance measures (namely, Euclidean distance, Pearson correlation and Spearman correlation) (41).
- SOM, an algorithm first proposed by Ritter and Kohonen for Artificial Neural Network, organizes features into spatially organized representations (42,43). We carried out the computation by the som() function from the R package som.
- MNM, proposed by Fraley and Raftery, fits a mixture of multivariate normal model to the data (44). We carried out the computation by the *Mclust( )* function from the R package *mclust* (45).

Clustering performance was assessed using the Adjusted Rand Index (ARI) in comparison with the ground truth (that is, the tumor subtype) (46–48). It ranges from 0 (when the clustering is essentially random) to 1 (when the clustering agrees perfectly with the ground truth).

All simulations were done using R version 4.1.0.

RESULTS

Clustering of data consisting of biological effects only

Results when there were 10% differentially expressed markers. When the amplification constant was as great as 2.4, all clustering methods except MNM performed perfectly or nearly perfectly (ARI ranging between 0.94 and 1.00), regardless of the normalization method used (Tables 1A and 1B). MNM performed well only when quantile normalization or VSN and/or ComBat were used (ARI: 0.94–0.98); it performed poorly when no normalization (ARI: 0.01) or median normalization (ARI: 0.68) was used without ComBat.

When the amplification constant lowered to 1.6, the results stayed similar (Tables 2A and 2B). All clustering methods except MNM and PAM Spearman performed perfectly or nearly perfectly (ARI: 0.92–1.00), regardless of the use of normalization or ComBat. PAM Spearman performed better than MNM but worse than the other clustering methods when ComBat was used (ARI: 0.80–0.92). MNM performed fairly well when quantile normalization and/or ComBat were used (ARI: 0.76–0.88), and poorly when no normalization (ARI: 0.01), median normalization (ARI: 0.00) or VSN (ARI: 0.00) was used without ComBat.

When the amplification constant decreased to 0.8, clustering accuracy deteriorated across the board, with *K*-means (when no or median normalization and no ComBat were used) being the most affected (ARI having decreased from 0.98 to <0.02) and PAM Pearson (regardless of normalization) the least (ARI having decreased from 1.00 to 0.94–0.98 when no ComBat was used and from 0.96–1.00 to 0.92–0.96 when ComBat was used) (Tables 3A and 3B). In this setting, MNM remained the worst clustering method regardless of the use of normalization or ComBat (ARI: <0.02). Quantile normalization was the best normalization method (ARI: 0.66–0.82 for PAM Spearman and 0.96–0.98 for *K*-means, Sparse *K*-means, PAM Euclidean and PAM Pearson), while VSN and median normalization had mixed performance depending on the clustering method

Table 3A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 0.8) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	0.01	0.02	0.98	0.98
Sparse *K*-means	0.48	0.50	0.96	1.00
PAM Euclidean	0.64	0.86	0.98	0.88
PAM Pearson	0.94	0.96	0.98	0.96
PAM Spearman	0.75	0.88	0.82	0.54
SOM	0.02	0.07	0.71	0.51
MNM	–0.00	0.00	0.01	0.01

Table 3B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 0.8) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	0.09	0.98	0.98	0.98
Sparse *K*-means	0.86	0.92	0.96	1.00
PAM Euclidean	0.73	0.96	0.96	0.94
PAM Pearson	0.96	0.92	0.96	0.92
PAM Spearman	0.69	0.73	0.66	0.05
SOM	0.61	0.64	0.96	0.88
MNM	0.00	–0.00	0.02	0.00

Table 4A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 0) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	–0.01	0.00	0.01	0.01
Sparse *K*-means	0.00	0.00	0.01	0.02
PAM Euclidean	0.25	–0.00	0.31	0.24
PAM Pearson	0.02	0.02	0.02	0.01
PAM Spearman	0.02	0.03	0.03	0.02
SOM	–0.00	0.00	0.00	0.00
MNM	0.00	0.00	0.02	0.01

Table 4B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (10%, 0) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	0.02	0.02	0.02	0.02
Sparse *K*-means	0.02	0.02	0.03	0.02
PAM Euclidean	0.39	0.47	0.46	0.39
PAM Pearson	0.03	0.51	0.02	0.03
PAM Spearman	0.37	0.03	0.03	0.01
SOM	0.00	0.00	0.04	0.01
MNM	0.00	–0.00	0.00	0.00

used. In particular, median normalization improved the clustering accuracy for PAM Spearman (ARI: 0.88 before ComBat and 0.73 after) but VSN worsened it (ARI: 0.54 and 0.05), compared with no normalization (ARI: 0.75 and 0.69).

When biological effects were not amplified, clustering accuracy plunged for all combinations of methods for normalization, BEC and clustering (ARI: <0.03) except PAM Euclidean with no ComBat and selected normalization methods (ARI: 0.24–0.31), PAM Euclidean with ComBat regardless of normalization (ARI: 0.39–0.47), PAM Pearson with ComBat and median normalization (ARI: 0.51), and PAM Spearman with ComBat and no normalization (ARI: 0.37) (Tables 4A and 4B).

Table 5A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 2.4) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	1.00	1.00	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	1.00	1.00	1.00	1.00
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	1.00	1.00	1.00	1.00
SOM	1.00	1.00	1.00	1.00
MNM	1.00	1.00	0.98	1.00

Table 5B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 2.4) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	1.00	1.00	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	1.00	1.00	1.00	1.00
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	1.00	1.00	1.00	1.00
SOM	1.00	1.00	1.00	1.00
MNM	1.00	1.00	0.98	1.00

Results when there were 30% differentially expressed markers. Increasing the percentage of differentially expressed markers improved the clustering accuracy. When the amplification constant was 2.4, all clustering methods except MNM performed perfectly (ARI: 1.00), regardless of the use of normalization or ComBat; MNM performed nearly perfectly to perfectly (ARI: 0.98–1.00) (Tables 5A and 5B). When the amplification constant was 1.6, clustering accuracy lessened only slightly (ARI: 0.84 for MNM with median normalization and without ComBat and 0.96–1.00 for the rest) (Tables 6A and 6B). When the amplification constant was 0.8, clustering accuracy diminished further to an ARI > 0.80 for all methods with a few exceptions when ComBat was not used: Sparse *K*-means with no normalization (ARI: 0.48) or median normalization (ARI: 0.50) and MNM with no normalization (ARI: 0.04) (Tables 7A and 7B). When biological effects were not amplified, all methods performed poorly except selected uses of *K*-means (with ComBat plus quantile normalization), PAM Euclidean (with ComBat and/or normalization), and PAM Pearson (with ComBat or VSN) (ARI: 0.36–0.78) (Tables 8A and 8B).

Clustering of data consisting of both biological effects and handling effects

Results when there were 10% differentially expressed markers. Supplementary Tables SX1 and SX7 show the clustering accuracy measured by the ARI averaged across the simulation runs, when the proportion of differential expression was 10% and the amplification constant was 2.4 for biological effects and 1 for handling effects, that is (π, c, d) equaled

Table 6A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 1.6) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	0.98	0.98	1.00	1.00
Sparse *K*-means	1.00	1.00	0.98	1.00
PAM Euclidean	0.98	1.00	1.00	1.00
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	1.00	1.00	1.00	1.00
SOM	0.98	0.98	1.00	1.00
MNM	0.98	0.84	0.98	0.98

Table 6B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 1.6) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	1.00	1.00	1.00	1.00
Sparse *K*-means	1.00	1.00	1.00	1.00
PAM Euclidean	0.98	0.98	1.00	1.00
PAM Pearson	1.00	1.00	1.00	1.00
PAM Spearman	1.00	0.98	0.96	1.00
SOM	0.98	0.98	1.00	1.00
MNM	0.98	0.98	0.98	1.00

Table 7A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 0.8) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	0.98	0.98	0.98	0.98
Sparse *K*-means	0.48	0.50	0.96	0.98
PAM Euclidean	0.98	0.94	0.96	0.96
PAM Pearson	0.96	0.96	0.96	0.98
PAM Spearman	0.96	0.96	0.96	0.96
SOM	0.80	0.90	0.98	0.98
MNM	0.04	0.82	0.84	0.90

Table 7B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 0.8) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	0.98	0.98	0.98	0.98
Sparse *K*-means	0.88	0.92	0.96	1.00
PAM Euclidean	0.86	0.92	0.98	0.98
PAM Pearson	0.98	0.90	0.98	0.98
PAM Spearman	0.96	0.96	0.88	0.94
SOM	0.98	0.98	0.98	0.98
MNM	0.82	0.82	0.84	0.94

Table 8A. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 0) in data with only biological effects. No ComBat was used

	None	Median	Quantile	VSN
K-means	0.01	0.01	0.68	0.71
Sparse *K*-means	0.21	0.23	0.01	0.02
PAM Euclidean	0.39	0.53	0.82	0.75
PAM Pearson	0.02	0.02	0.05	0.47
PAM Spearman	0.04	0.05	0.05	0.05
SOM	0.01	0.04	0.09	0.13
MNM	0.02	0.01	0.02	0.02

Table 8B. The clustering accuracy (measured by the Adjusted Rand Index, ARI) when (π, c) equaled (30%, 0) in data with only biological effects. ComBat (with array slides as the batch variable) was used after normalization

	None	Median	Quantile	VSN
K-means	0.09	0.22	0.78	0.78
Sparse *K*-means	0.02	0.02	0.02	0.03
PAM Euclidean	0.40	0.37	0.43	0.62
PAM Pearson	0.56	0.03	0.71	0.62
PAM Spearman	0.41	0.04	0.05	0.05
SOM	0.14	0.17	0.22	0.21
MNM	0.03	0.01	0.02	0.02

(10%, 2.4, 1). Figure 1 shows the distribution of the ARI as the median and the inter-quartile range (IQR).

- When ComBat was not used while biological effects and handling effects were partially confounded, most of the clustering methods performed very well with or without dependence on normalization (Figure 1, panel A). Three clustering methods (PAM Pearson, PAM Euclidean and Sparse *K*-means) performed perfectly or nearly perfectly (mean ARI ranging from 0.97 to 1.00), with or without normalization. Another three clustering methods (PAM Spearman, *K*-means and SOM) performed well only with normalization regardless of the method used (mean ARI: 0.09–0.86 before normalization and 0.81–1.00 after normalization). MNM, relatively the worst clustering method, performed well when quantile normalization was used (mean ARI: 0.94), moderately well for median normalization (mean ARI: 0.62), and poorly for VSN (mean ARI: 0.27) or no normalization (mean ARI: 0.02).
- While biological effects and handling effects were partially confounded, the use of ComBat led to plunged performance across the board (Figure 1, panel B). PAM Pearson was the only method that performed moderately well without normalization (mean ARI: 0.75). Its performance worsened with the use of normalization to various degrees (mean ARI: 0.66 for quantile normalization, 0.59 for median normalization, and 0.29 for VSN). Although Sparse *K*-means performed fairly poorly before normalization (mean ARI: 0.30), its clustering accuracy was improved significantly by quantile normalization (mean ARI: 0.76) and moderately by median normalization (mean ARI: 0.31) but worsened by VSN (mean ARI: 0.18). Similarly, PAM Euclidean performed poorly before normalization (mean ARI: 0.16), which was improved significantly by quantile normalization (mean ARI: 0.71), moderately by VSN (mean ARI: 0.28), and slightly by median normalization (mean ARI: 0.20). *K*-means and SOM performed poorly (mean ARI: 0.02 and 0.01 for no normalization, 0.05 and 0.02 for median normalization, and 0.01 and 0.09 for VSN), with the exception of quantile normalization (mean ARI: 0.79 and 0.74). MNM performed extremely badly regardless of normalization (mean ARI: 0.00–0.02). The reason for the negative impact of ComBat in this scenario may be that biological signals between clusters have been unintentionally removed when adjusting the data with a Com-

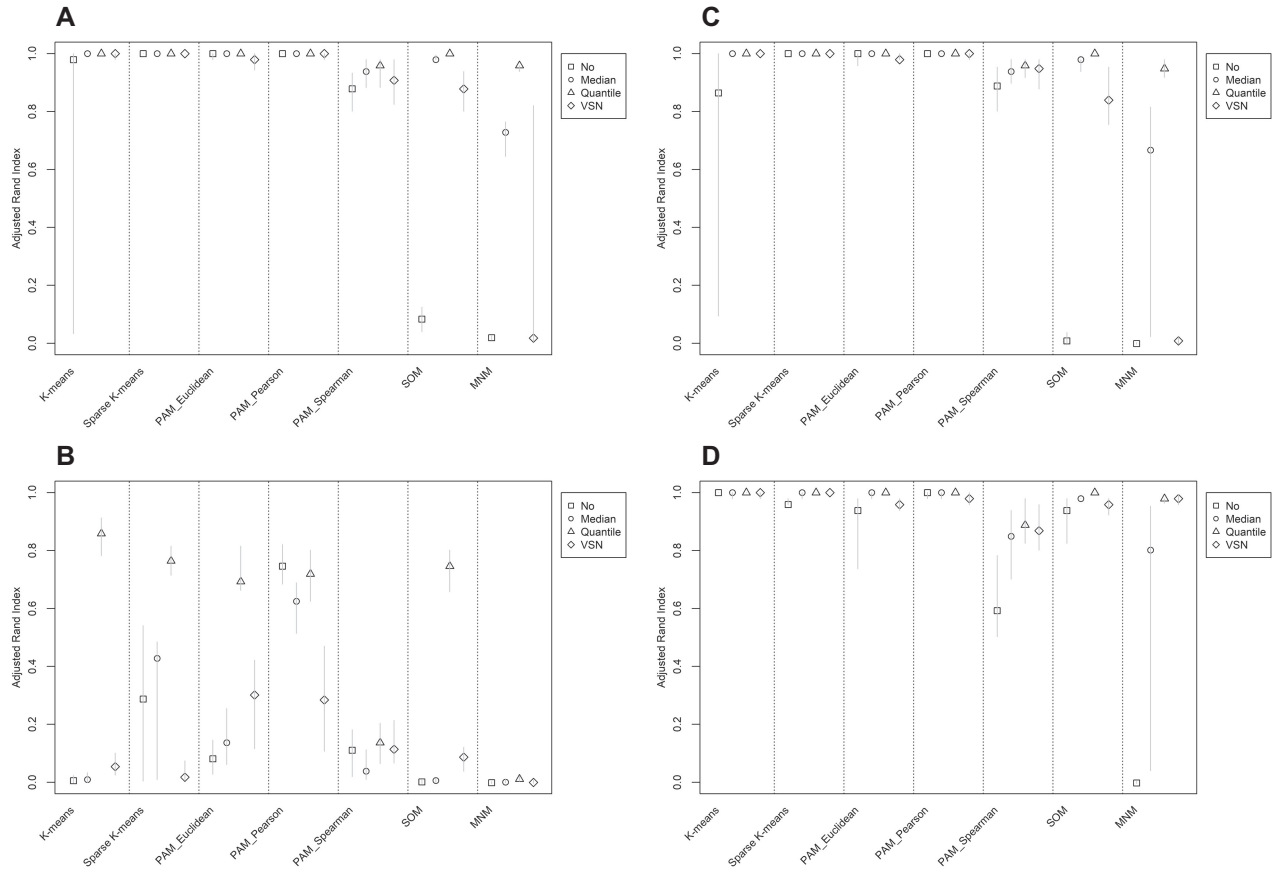

Figure 1. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 10% and the amplification constant was 2.4 for biological effects and 1 for handling effects, that is (π, c, d) equaled (10%, 2.4, 1). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

Bat regression model that includes not a cluster variable (as it is not known at the time of data preprocessing) but only a batch variable that confounds with the underlying clustering (Supplementary Figures SW1–SW6).

- When biological effects and handling effects were balanced via stratification, regardless of the use of ComBat, clustering methods and normalization methods performed similarly well to the scenario with confounding and no ComBat (Figure 1, panels C and D). Here the use of ComBat not only did not harm but rather improved the performance of selected methods that previously performed poorly (that is, SOM when combined with no normalization or VSN, and MNM when combined with VSN). In other words, the benefit of balanced biological effects and handling effects was mainly in providing immunity to potential negative impacts of ComBat.
- Across all four panels and all seven clustering methods, quantile normalization was consistently the best performer compared with median normalization and VSN.

Supplementary Tables SX2 and SX8 present the mean ARI when (π, c, d) was (10%, 1.6, 1); Figure 2 presents the median ARI and the IQR. With a smaller increase of the magnitude of differential expression, as expected, the clustering accuracy decreased for some of the methods. Notably, *K*-means, SOM and MNM performed well only with quantile normalization (mean ARI: 0.85–0.98) and quite poorly otherwise (mean ARI: 0.01–0.41); PAM Spearman worsened moderately across normalization methods (mean ARI: 0.63–0.82); PAM Euclidean worsened slightly when used with no normalization (mean ARI: 0.74) or VSN (mean ARI: 0.87). Nevertheless, the relative performance of clustering methods and normalization methods and the impact of ComBat and balanced design remained similar to those in Figure 1. That is, PAM Pearson was the best clustering method regardless of the use of normalization; quantile normalization was the best performer and for selected clustering methods the only well-performer; ComBat was harmful when biological effects and handling effects were confounded; balanced design brought little benefit except immunity to the harm from ComBat.

Supplementary Tables SX3 and SX9 display the mean ARI when (π, c, d) was (10%, 0.8, 1); Figure 3 displays the median ARI and the IQR. With an even smaller increase of the magnitude of differential expression, clustering accu-

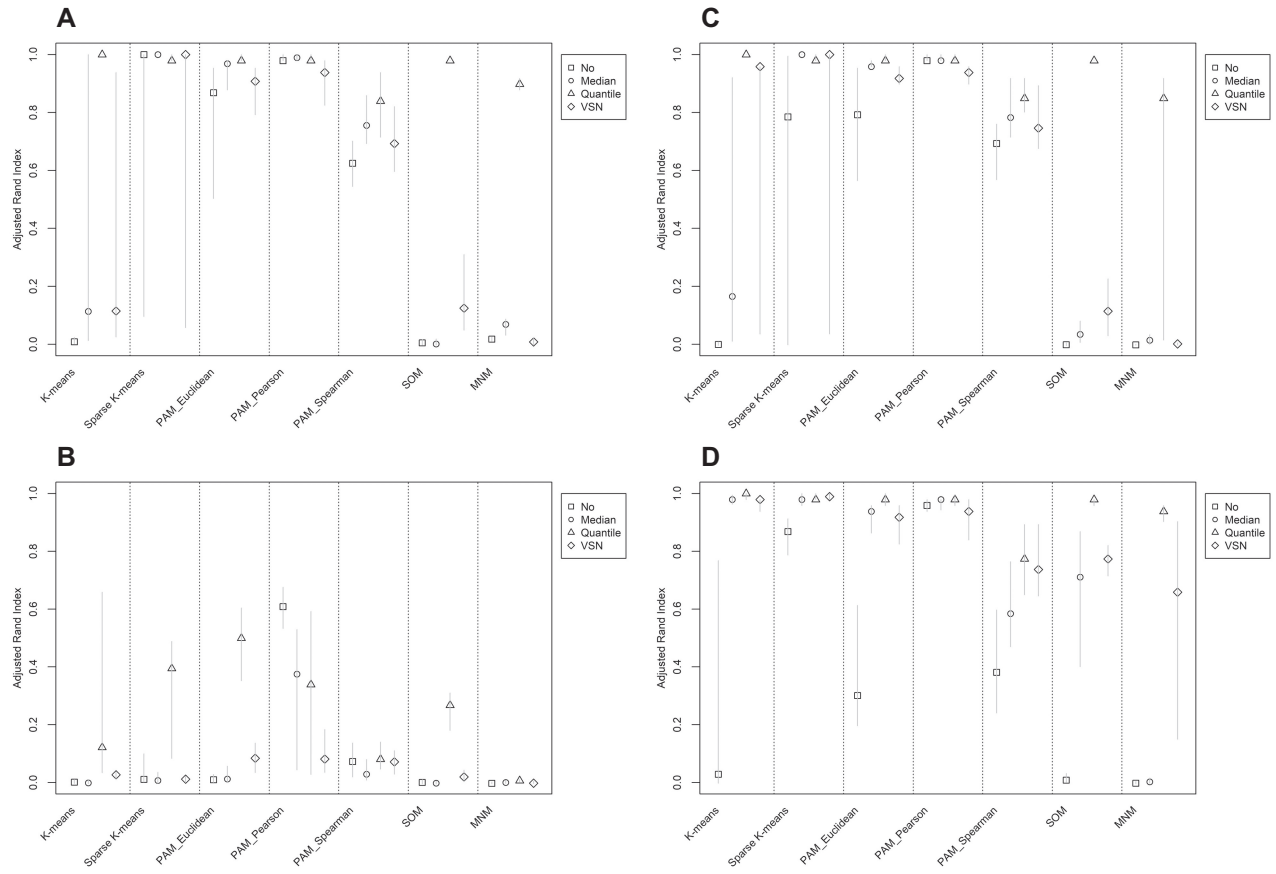

Figure 2. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 10% and the amplification constant was 1.6 for biological effects and 1 for handling effects, that is (π, c, d) equaled (10%, 1.6, 1). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

racy worsened for more methods and to a greater extent. For example, in addition to *K*-means, SOM and MNM, Sparse *K*-means also performed well only with quantile normalization (mean ARI: 0.89) and quite poorly otherwise (mean ARI: 0.03–0.22); PAM Euclidean performed reasonably well only with quantile normalization (mean ARI: 0.88) and moderately well to poorly otherwise (mean ARI: 0.58 for VSN, 0.36 for median normalization, and 0.13 for no normalization); MNM performed very poorly across all normalization methods (mean ARI: 0.01–0.02). Again, the relative performance of clustering methods and normalization methods and the impact of ComBat and balanced design remained similar to those in Figure 1.

Supplementary Tables SX4 and SX10 render the mean ARI when (π, c, d) was (10%, 1.6, 3); Figure 4 renders the median ARI and the IQR. With the amplified handling artifacts, the performance of all normalization and clustering methods deteriorated badly (ARI: near 0), except for (i) two methods (PAM Euclidean and PAM Pearson) when they were used with quantile normalization and without ComBat (mean ARI: 0.45–0.86) for data with confounding handling effects and (ii) five methods (*K*-means, Sparse *K*-means, PAM Euclidean, PAM Pearson and SOM) when used with quantile normalization plus ComBat for balanced data (ARI: 0.81–0.95). Here balanced design, when in combination with quantile normalization and ComBat, offered the benefit of effectuating more clustering methods.

Results when there were 30% differentially expressed markers. Supplementary Tables SX5 and SX11 depict the mean ARI when (π, c, d) was (30%, 1.6, 1); Figure 5 depicts the median ARI and the IQR. When biological signals were enhanced by increasing the proportion of differential expression (through removing a portion of the non-differentially expressed markers), similar findings were seen as to those when increasing the magnitude of mean differences among the differentially expressed markers with (π, c, d) being (10%, 2.4, 1). Limited differences included the improved performance of MNM with VSN and no ComBat and the worsened performance of all clustering methods when ComBat was used for data with confounding handling effects.

Supplementary Tables SX6 and SX12 convey the mean ARI when (π, c, d) was (30%, 0.8, 1); Figure 6 conveys the median ARI and the IQR. The clustering accuracy in this

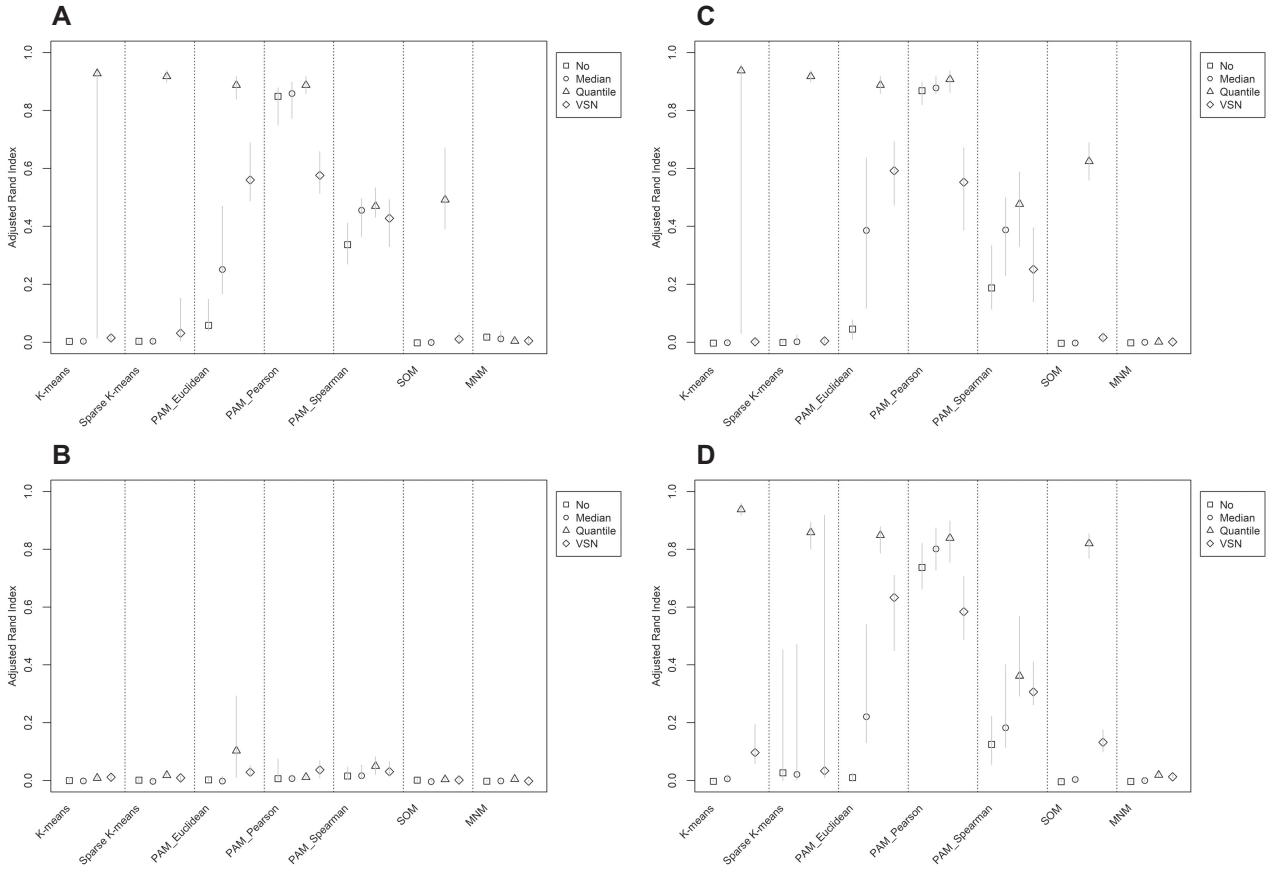

Figure 3. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 10% and the amplification constant was 0.8 for biological effects and 1 for handling effects, that is (π, c, d) equaled (10%, 0.8, 1). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

scenario was in between that for ARI being (10%, 1.6, 1) and (30%, 1.6, 1). Once again, the relative performance of clustering methods and normalization methods and the impact of ComBat and balanced design remained similar to those in Figures 1–3 and 5.

In addition to the above results when Combat was not used or used (with array slides as the batch variable) after normalization, we also examined when the order of ComBat and normalization changed and when the choice of the batch variable became experimental batches. We found that their results stayed very similar and that it was slightly more effective to use array slides for defining the batch variable in ComBat (Supplementary Figures SX1–SX6 and Tables SX1–SX12).

Clustering when the two clusters had unequal sample sizes

To evaluate the robustness of our findings when the two clusters had unequal sample sizes, we conducted further simulations when one cluster kept the 96 sample and the other had 48 samples randomly selected out of the 96 samples. In these simulations, we observed similar performance of normalization and ComBat for the four better-performing clustering methods (namely, PAM Pearson, PAM Euclidean, *K*-means and Sparse *K*-means), in the range of signal to noise ratios examined in our study, compared with the results when the two clusters had equal sample sizes; in addition, we saw much worse performance of some to all normalization methods for the three worse-performing clustering methods (namely, PAM Spearman, SOM and MNM). Results for these simulations are presented in Supplementary Figures SY1-SY6 and Supplementary Tables SY1-SY6.

Take the scenario of (π, c, d) of (10%, 2.4, 1) as an example (Supplementary Figure SY1). When biological effects and handling effects were partially confounded while ComBat was not used (panel A), or when biological effects and handling effects were balanced (panels C and D), PAM Pearson, PAM Euclidean and Sparse *K*-means performed perfectly or nearly perfectly with or without normalization (mean ARI: 0.65–1.00); *K*-means performed well if median or quantile normalization was used (mean ARI: 0.87–1.00); PAM Spearman and MNM performed badly despite the use of normalization (mean ARI: 0.01–0.40). When biological effects and handling effects were partially confounded while ComBat was used (panel B), all

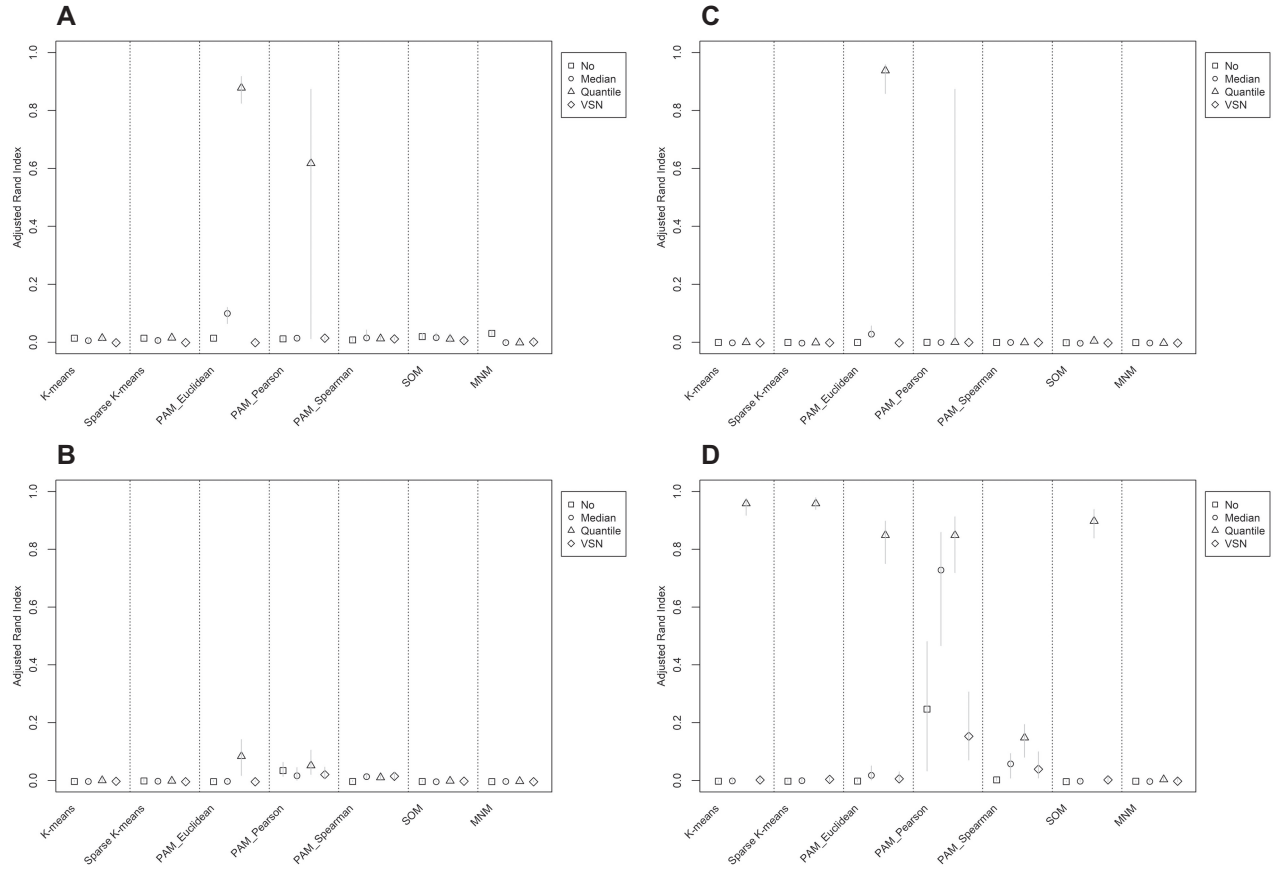

Figure 4. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 10% and the amplification constant was 1.6 for biological effects and 3 for handling effects, that is (π, c, d) equaled (10%, 1.6, 3). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

combinations of clustering and normalization methods performed badly (mean ARI: 0.00–0.29), except quantile normalization (when not used with PAM Spearman or MNM) (mean ARI: 0.66–0.89) and PAM Pearson (when not used with VSN) (mean ARI: 0.53–0.71).

Clustering when the number of clusters was mis-specified

In practice, the number of clusters is often unknown a priori and can be mis-specified when applying a clustering algorithm. As such we conducted additional simulations when the number of clusters was mis-specified to be three while the underlying number of clusters was two. Here we observed similar relative performance but moderately reduced absolute performance across the methods for normalization, BEC and clustering in the six signal-to-noise ratio scenarios we examined. Results for these simulations are presented in Supplementary Materials (Supplementary Figures SZ1–SZ6 and Supplementary Tables SZ1-SZ6). Generally speaking, PAM Pearson remained the best clustering method and quantile normalization the best normalization method; the use of ComBat was still detrimental when handling effects were confounding with biological effects; PAM Spearman and MNM again tended to be the worst performing clustering methods and VSN the worst performing normalization method.

DISCUSSION

Large-scale genomic studies such as The Cancer Genome Atlas have provided a deeper understanding of the molecular alterations involved in carcinogenesis and confirmed the view that cancer represents a wide variety of diseases that can be divided into molecular subtypes (5,49). The discovery of new molecular subtypes has been shown to be influenced by data normalization in the context of two-color cDNA arrays (20). Here, we studied this issue in miRNA arrays using *in silico* data that were generated by re-sampling from a pair of data sets to mimic real data distribution characteristics and create a range of scenarios of differential expression signals and experimental handling artifacts. This simulation algorithm makes the additivity assumption, which has been deemed reasonable for microarray data and has been adopted in published methods for microarray data normalization and analysis (50,51).

Our study showed that the effectiveness of discovering sample clusters depends extensively on the choice of method

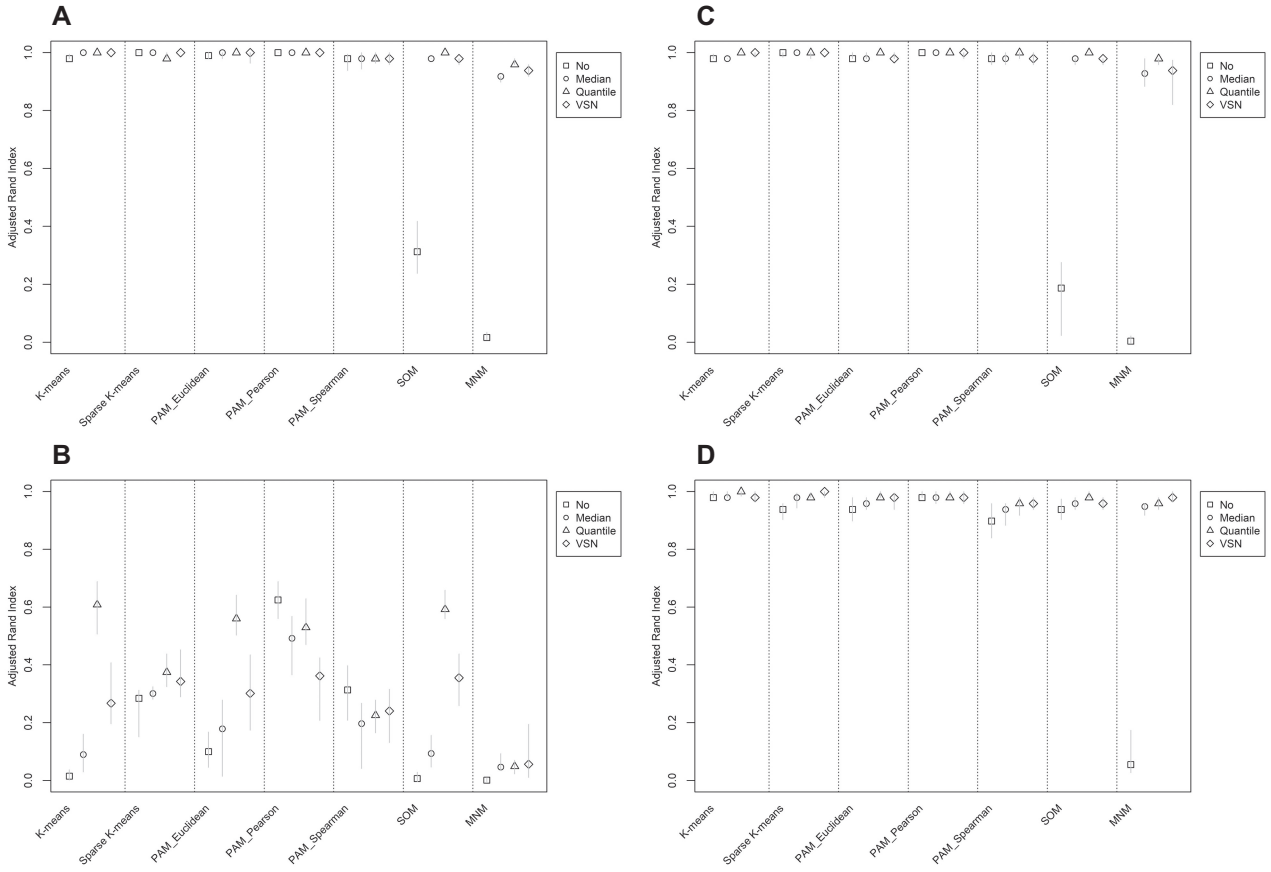

Figure 5. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 30% and the amplification constant was 1.6 for biological effects and 1 for handling effects, that is (π, c, d) equaled (30%, 1.6, 1). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

for sample clustering, data normalization and BEC, the latter of which further depends on the study design for the sample assignment process. Among the methods we examined, quantile normalization was the best performer for data normalization, while VSN was largely the worst; PAM Pearson was the best method for clustering, while PAM Spearman and MNM were the worst; the use of BEC (that is, ComBat in our study) was detrimental for accurately discovering the subgrouping structure when handling effects and biological signals were confounded, and it brought very limited benefits when they were balanced. When applying ComBat, it was slightly more effective to use array slides for defining batches. Of note, to make the clustering more objective and comparable across methods, no extra feature selection step was taken except what was allowed by the tuning parameter in Sparse *K*-means; addition of such a step may improve the performance of certain clustering methods such as MNM, which warrants further research and is beyond the scope of this article.

In our previous studies, we found that normalization can lead to biased estimation of the classification error rate in the optimistic direction for cross-validation and that quantile normalization is detrimental for identifying prognostic biomarkers and building outcome predictors. Unlike these findings, we found in this study that normalization is generally beneficial for sample clustering and quantile normalization is the best performer. This finding is similar to how normalization behaves for differential expression analysis, where it helps remove the bias in the estimation of group means for each marker and hence the distributional separation of the two sample clusters (Supplementary Figures SW1–SW6) (52). Our findings are robust to whether the clusters are of comparable sizes and whether the number of clusters is correctly specified when applying the clustering algorithms.

In conclusion, our study shed lights on the operating characteristics of normalization and BEC for the discovery of tumor subtypes and encourage the use of quantile normalization in combination with a well-performing clustering method.

DATA AVAILABILITY

Human tumor tissues used in this study were obtained from participants who provided informed consent, and their use in our study was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board. The R packages

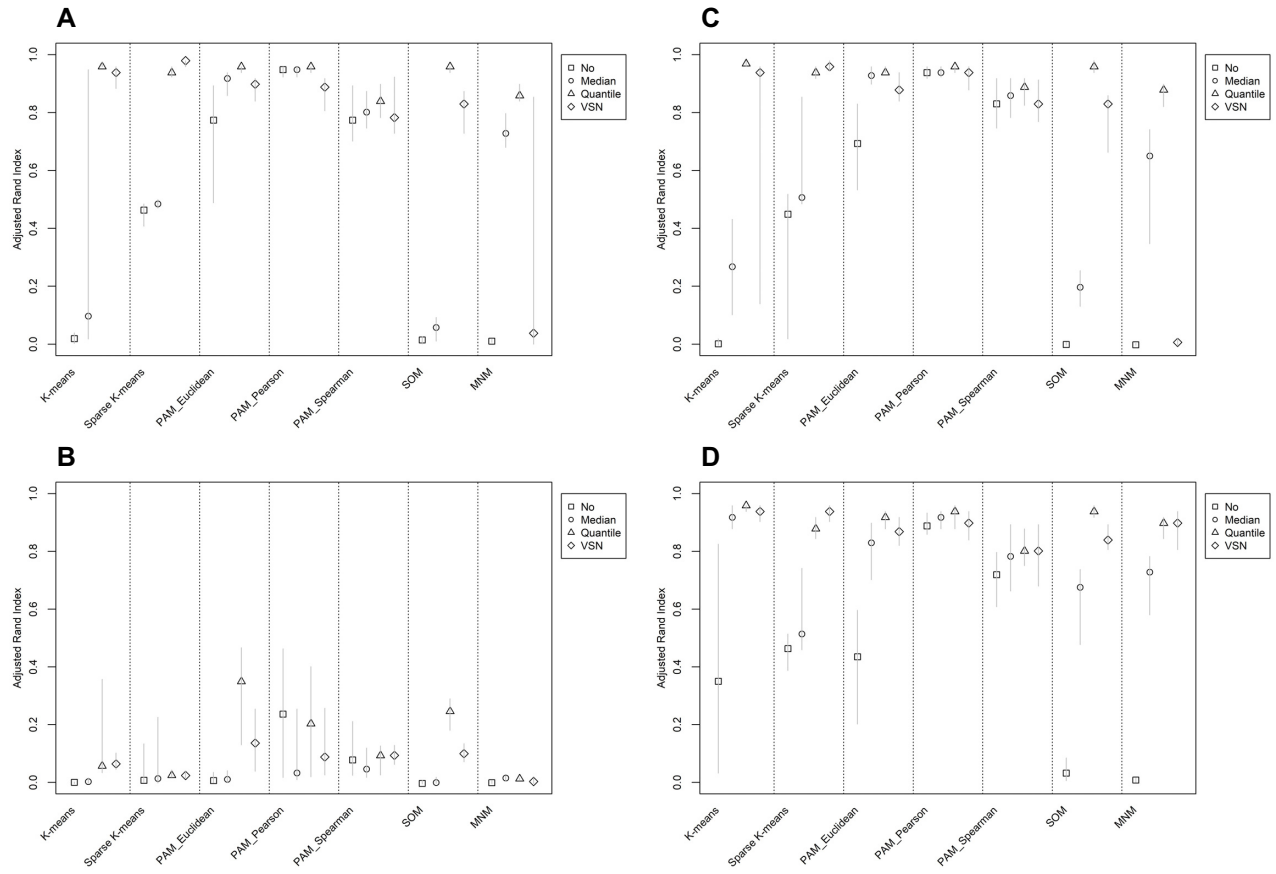

Figure 6. Plots of the clustering accuracy (measured by the Adjusted Rand Index, ARI) when the proportion of differential expression was 30% and the amplification constant was 0.8 for biological effects and 1 for handling effects, that is (π, c, d) equaled (30%, 0.8, 1). The inter-quartile range of the ARI among 30 simulation runs are represented by a vertical bar and the median by the symbol in each bar. Left panels show the results when there was *partial confounding* between handling effects and biological effects; right panels show the results when there was *balance via stratification* between biological effects and handling effects. Top panels show the results *without* ComBat; bottom panels show results *with* ComBat (with array slides as the batch variable) after normalization.

containing the data and code used for the simulations can be downloaded at https://doi.org/10.5281/zenodo.7314352 and https://doi.org/10.5281/zenodo.7314358, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

National Institutes of Health [CA214845, CA217694, HG012124, CA008748 to L.X.Q.]. Funding for open access charge: NIH [HG012124].
Conflict of interest statement. None declared.

REFERENCES

1. Malone,E.R., Oliva,M., Sabatini,P.J.B., Stockley,T.L. and Siu,L.L. (2020) Molecular profiling for precision cancer therapies. *Genome Med.*, **12**, 8.
2. Liu,D., Augello,M.A., Grbesa,I., Prandi,D., Liu,Y., Shoag,J.E., Karnes,R.J., Trock,B.J., Klein,E.A., Den,R.B. *et al.* (2021) Tumor subtype defines distinct pathways of molecular and clinical progression in primary prostate cancer. *J. Clin. Invest.*, **131**, e147878.
3. Perou,C.M., Sørlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.
4. Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 10869–10874.
5. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J., Ding,L., Mardis,E.R. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
6. Natrajan,R. and Weigelt,B. (2016) Risk stratification and intrinsic subtype classification of breast cancer: a multi-Parameter test to rule them all?*J. Natl. Cancer Inst.*, **108**, djw118.
7. Weigelt,B. and Reis-Filho,J.S. (2010) Molecular profiling currently offers no more than tumour morphology and basic immunohistochemistry. *Breast Cancer Res.*, **12**(Suppl. 4), S5.
8. Garge,N.R., Page,G.P., Sprague,A.P., Gorman,B.S. and Allison,D.B. (2005) Reproducible clusters from microarray research: whither? *BMC Bioinf.*, **6**(Suppl. 2), S10.
9. Patil,P., Bachant-Winner,P.O., Haibe-Kains,B. and Leek,J.T. (2015) Test set bias affects reproducibility of gene signatures. *Bioinformatics*, **31**, 2318–2323.
10. Elloumi,F., Hu,Z., Li,Y., Parker,J.S., Gulley,M.L., Amos,K.D. and Troester,M.A. (2011) Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med. Genomics*, **4**, 54.

11. Leek,J.T., Scharpf,R.B., Bravo,H.C., Simcha,D., Langmead,B., Johnson,W.E., Geman,D., Baggerly,K. and Irizarry,R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.
12. Kensler,K.H., Sankar,V.N., Wang,J., Zhang,X., Rubadue,C.A., Baker,G.M., Parker,J.S., Hoadley,K.A., Stancu,A.L., Pyle,M.E. *et al.* (2019) PAM50 Molecular intrinsic subtypes in the nurses' Health study cohorts. *Cancer Epidemiol. Biomarkers Prev.*, **28**, 798–806.
13. Peixoto,L., Risso,D., Poplawski,S.G., Wimmer,M.E., Speed,T.P., Wood,M.A. and Abel,T. (2015) How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.*, **43**, 7664–7674.
14. Lusa,L., McShane,L.M., Reid,J.F., De Cecco,L., Ambrogi,F., Biganzoli,E., Gariboldi,M. and Pierotti,M.A. (2007) Challenges in projecting clustering results across gene expression-profiling datasets. *J. Natl. Cancer Inst.*, **99**, 1715–1723.
15. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
16. Ni,A. and Qin,L.X. (2021) Performance evaluation of transcriptomics data normalization for survival risk prediction. *Brief. Bioinform.*, **22**, bbab2575 .
17. Wu,Y., Huang,H.C. and Qin,L.X. (2021) Making external validation valid for molecular classifier development. *JCO Precis. Oncol.*, **5**, 1250–1258.
18. Huang,H.C. and Qin,L.X. (2018) Empirical evaluation of data normalization methods for molecular classification. *PeerJ*, **6**, e4584.
19. Qin,L.X., Huang,H.C. and Begg,C.B. (2016) Cautionary note on using cross-Validation for molecular classification. *J. Clin. Oncol.*, **34**, 3931–3938.
20. Freyhult,E., Landfors,M., Önskog,J., Hvidsten,T.R. and Rydén,P. (2010) Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC Bioinf.*, **11**, 503.
21. He,L., Thomson,J.M., Hemann,M.T., Hernando-Monge,E., Mu,D., Goodson,S., Powers,S., Cordon-Cardo,C., Lowe,S.W., Hannon,G.J. *et al.* (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828–833.
22. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
23. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
24. Qin,L.X., Zhou,Q., Bogomolniy,F., Villafania,L., Olvera,N., Cavatore,M., Satagopan,J.M., Begg,C.B. and Levine,D.A. (2014) Blocking and randomization to improve molecular biomarker discovery. *Clin. Cancer Res.*, **20**, 3371–3378.
25. Qin,L.X., Huang,H.C., Villafania,L., Cavatore,M., Olvera,N. and Levine,D.A. (2018) A pair of datasets for microRNA expression profiling to examine the use of careful study design for assigning arrays to samples. *Sci. Data*, **5**, 180084.
26. Qin,L.X. and Levine,D.A. (2016) Study design and data analysis considerations for the discovery of prognostic molecular biomarkers: a case study of progression free survival in advanced serous ovarian cancer. *BMC Med. Genomics*, **9**, 27.
27. Qin,L.X., Huang,H.C. and Zhou,Q. (2014) Preprocessing steps for agilent MicroRNA arrays: does the order matter?*Cancer Inform.*, **13**, 105–109.
28. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
29. Chawade,A., Alexandersson,E. and Levander,F. (2014) Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *J. Proteome Res.*, **13**, 3114–3120.
30. Välikangas,T., Suomi,T. and Elo,L.L. (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform*, **19**, 1–11.
31. Rao,Y., Lee,Y., Jarjoura,D., Ruppert,A.S., Liu,C.G., Hsu,J.C. and Hagan,J.P. (2008) A comparison of normalization techniques for microRNA microarray data. *Stat. Appl. Genet. Mol. Biol.*, **7**, Article22.
32. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**(Suppl), 496–501.
33. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
34. Bolstad,B.M. (2013) preprocessCore: a collection of pre-processing functions. R package version 1.
35. Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**(Suppl. 1), S96–S104.
36. Forgy,E.W. (1965) Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometrics*, **21**, 768–769.
37. Lloyd,S. (1982) Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, **28**, 129–137.
38. Witten,D.M. and Tibshirani,R. (2010) A framework for feature selection in clustering. *J. Am. Stat. Assoc.*, **105**, 713–726.
39. Witten,D.M. and Tibshirani,R. (2013) sparcl: perform sparse hierarchical clustering and sparse k-means clustering. R package version 1.
40. Kaufman,L. and Rousseeuw,P.J. (2009) In: *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
41. Maechler,M., Rousseeuw,P., Struyf,A., Hubert,M. and Hornik,K. (2012) Cluster: cluster analysis basics and extensions. R package version 1.56.
42. Kohonen,T. (1990) The self-organizing map. *Proc. IEEE*, **78**, 1464–1480.
43. Ritter,H. and Kohonen,T. (1989) Self-organizing semantic maps. *Biol. Cybern.*, **61**, 241–254.
44. Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Assoc.*, **97**, 611–631.
45. Scrucca,L., Fop,M., Murphy,T.B. and Raftery,A.E. (2016) mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R Journal*, **8**, 289.
46. Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Assoc.*, **66**, 846–850.
47. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.
48. Vinh,N.X., Epps,J. and Bailey,J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, **11**, 2837–2854.
49. Cancer Genome Atlas Research, N. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
50. Qin,L.X. and Satagopan,J.M. (2009) Normalization method for transcriptional studies of heterogeneous samples–simultaneous array normalization and identification of equivalent expression.*Stat. Appl. Genet. Mol. Biol.*, **8**, Article 10.
51. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data.*J. Comput. Biol.*, **7**, 819–837.
52. Qin,L.X. and Zhou,Q. (2014) MicroRNA array normalization: an evaluation using a randomized dataset as the benchmark. *PLoS One*, **9**, e98879.