Published in final edited form as:

Stat Med. 2018 December 20; 37(29): 4386–4403. doi:10.1002/sim.7928.

Assisted gene expression-based clustering with AWNCut

Yang Li^{1,2}, Ruofan Bie², Sebastian J Teran Hidalgo⁵, Yichen Qin⁴, Mengyun Wu^{3,5}, and Shuangge Ma^{2,5}

¹Center for Applied Statistics, Renmin University of China, Beijing, China

²School of Statistics, Renmin University of China, Beijing, China

³School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

⁴Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, Cincinnati, Ohio

⁵Department of Biostatistics, Yale University, New Haven, Connecticut

Abstract

In the research on complex diseases, gene expression (GE) data have been extensively used for clustering samples. The clusters so generated can serve as the basis for disease subtype identification, risk stratification, and many other purposes. With the small sample sizes of genetic profiling studies and noisy nature of GE data, clustering analysis results are often unsatisfactory. In the most recent studies, a prominent trend is to conduct multidimensional profiling, which collects data on GEs and their regulators (copy number alterations, microRNAs, methylation, etc.) on the same subjects. With the regulation relationships, regulators contain important information on the properties of GEs. We develop a novel *assisted clustering* method, which effectively uses regulator information to improve clustering analysis using GE data. To account for the fact that not all GEs are informative, we propose a weighted strategy, where the weights are determined data-dependently and can discriminate informative GEs from noises. The proposed method is built on the NCut technique and effectively realized using a simulated annealing algorithm. Simulations demonstrate that it can well outperform multiple direct competitors. In the analysis of TCGA cutaneous melanoma and lung adenocarcinoma data, biologically sensible findings different from the alternatives are made.

Correspondence: Mengyun Wu, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China; or Department of Biostatistics, Yale University, New Haven, CT 06520-8034. mengyun.wu@yale.edu, Shuangge Ma, School of Statistics, Renmin University of China, Beijing 100872, China; or Department of Biostatistics, Yale University, New Haven, CT 06520-8034. shuangge.ma@yale.edu.

AUTHOR CONTRIBUTIONS

All authors were involved in methodological development. Ruofan Bie conducted programming and numerical study. All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

FINANCIAL DISCLOSURE

None reported.

Keywords

assisted analysis; clustering; gene expression data; NCut

1 | INTRODUCTION

For many complex diseases, gene expression (GE) profiling studies have been extensively conducted.^{1,2} Clustering has been widely conducted, searching for groups of subjects with distinct GE patterns.³ The results so generated have important implications for disease subtype identification, risk stratification, and many other purposes. Promising findings have been made for cancer, cardiovascular diseases, diabetes, and many others.⁴ Popular statistical techniques range from those based on the “classic” Kmeans and hierarchical clustering to the more recent model-based clustering, graph-based clustering, and others. For example, an enhanced deterministic Kmeans clustering algorithm has been developed for detecting cancer subtypes with GE data, involving a novel and systematic method for selecting initial centroids.⁵ In another study, a novel sparse framework for Kmeans and hierarchical clustering has been proposed for identifying important genes and subject subgroups simultaneously, using a Lasso-type penalty.⁶ Other relevant approaches include regularized Kmeans clustering,⁷ entropy-based consensus clustering,⁸ and convex clustering.⁹ There are also recent works on model-based clustering that usually describe clusters using their probability distributions with a mixture model.¹⁰ Representative examples include the expanded parsimonious Gaussian mixture model,¹¹ Bayesian model-based clustering,¹² and others. Nowadays, the graph-based clustering methods have become increasingly popular due to their lucid interpretations and satisfactory clustering accuracy. For example, approaches have been developed based on the normalized cut (NCut) algorithm for subtype discovery from GE data.^{13,14} Comprehensive discussions are available in the literature.^{15,16} With sample, cost, and other constraints, most of the existing GE studies have small sample sizes. On the other side, with the maturity of profiling techniques, usually, the expressions of a large number of genes are measured. For complex diseases, such as cancer, the signal-to-noise ratios are usually low. As such, clustering results based on GE data are still often unsatisfactory.

In the most recent biomedical research, a prominent trend is to conduct multidimensional studies, which collect measurements on GEs and their regulators, including copy number alterations (CNAs), microRNAs, methylation, and others, on the same subjects. A representative example is TCGA (The Cancer Genome Atlas), which is a collective effort organized by NCI/NIH of the USA and has conducted extensive profiling on multiple cancer types. We note that, in some studies, measurements at the downstream of GEs, for example, protein expressions, have also been collected. In this article, we focus on GEs and their upstream regulators, which are more commonly measured. The proposed analysis can potentially be extended to accommodate the downstream measurements of GEs. Regulators regulate GE levels and contain important information on the properties of GEs. In a series of recent studies, data on GEs and their regulators have been analyzed in an integrated manner and generated interesting findings beyond gene-expression-only analysis (for example, superior outcome models).¹⁷ Different from these studies, some other studies have the

objective of more accurately identifying GE markers and building predictive models for disease outcomes *with the assistance of* regulators. Examples include the assisted robust marker identification approach,¹⁸ collaborative regression,¹⁹ shared informative factor models,²⁰ canonical variate regression,²¹ and others. Besides the supervised regression models, there are also multidimensional studies that focus on unsupervised clustering. For example, Shen et al have proposed a penalized latent variable regression method to identify common latent variables for detecting disease subtypes with multiple omics datasets.²² A clustering procedure based on the Joint and Individual Variance Explained algorithm has been developed to find joint and data type-specific clusters simultaneously.²³ Another relevant study is the recent ANCut,²⁴ which aims at more effectively clustering GEs with the assistance of regulators. For more detailed discussions, we refer to the literature.^{25,26} In some of the aforementioned studies, the “central status” of GEs has been driven by the extensive data availability and their actionability in therapeutic development.

Motivated by the importance of clustering analysis, limitations of gene-expression-only analysis, and recent successes of assisted analysis in other contexts, in this study, our goal is to develop an assisted clustering method to identify distinct groups of subjects based on GE patterns with the assistance of regulators. Different from supervised marker selection and model building,^{18,19} our analysis is on unsupervised clustering. Although seemingly connected to the ANCut,²⁴ the present study is fundamentally different. The goal of ANCut²⁴ is to cluster GEs, and regulators are used to more accurately compute the “distance” between GEs. In contrast, in the present study, the goal is to cluster subjects based on both GEs and regulators. With a significantly different analysis goal, this study is warranted beyond the existing assisted analysis. Moreover, this study targets at addressing another important problem. With many clustering methods, for example, the “classic” Kmeans and hierarchical clustering, all GEs are used in an equal manner. When a large number of GEs are collected, many or even most of them are expected to be irrelevant, and only a small subset are informative for differentiating subjects. Including a large number of “noisy” GEs in clustering may mask true signals and lead to inferior results, which can be partly seen from our numerical analysis. Motivated by the successes of clustering methods that also have a selection component, for example, the sparse Kmeans,⁶ we seek to incorporate the selection (of relevant GEs and regulators) component in the proposed assisted analysis. The proposed novel developments are based on the NCut technique, which is originally developed in imaging and other scientific fields and more recently applied in genetic studies and shown to have competitive performance.²⁷ As such, this study will also contribute to methodological development of the novel NCut technique.

2 | METHODS

Consider a dataset with n independent subjects. For each subject, measurements are available on both GEs and regulators. Note that the proposed analysis does not assume known regulation relationships or require the collection of all relevant regulators and, hence, is sufficiently flexible. In Section 4, with TCGA data, we analyze CNAs, whose regulations of GEs have been well examined. Other types of regulators, such as microRNA, can be analyzed in the same manner. When there are multiple types of regulators, we can follow the recent literature,²⁸ where a “mega” regulator vector is created by stacking them together. For

the i th subject, let $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})$ and $\mathbf{Z}_i = (z_{i1}, \dots, z_{iq})$ be the p - and q -vectors of GE and CNA measurements. Denote $\mathbf{X}$ and $\mathbf{Z}$ as the $n \times p$ and $n \times q$ design matrices composed of $\mathbf{X}_i$ and $\mathbf{Z}_i$. For the completeness of this study, we first briefly review the NCut approach for clustering subjects based on GEs only, and more details are available in the literature.²⁹

2.1 | Clustering analysis with NCut

NCut is related to the formulation of grouping under graph theories, where an undirected graph is constructed with nodes representing subjects. For n subjects, an adjacency matrix $\mathbf{U} = (u_{il})_{n \times n}$ is computed based on this graph, where $u_{il} \geq 0$ measures the similarity between subjects i and l based on the GEs. There are multiple ways of defining u_{il} , such as the inverse of Euclidean distance $1/\sqrt{\sum_{j=1}^p (x_{ij} - x_{lj})^2}$ and inverse of divergence

$1/[\sum_{j=1}^p (x_{ij} \log(x_{ij}/x_{lj}) - x_{ij} + x_{lj})]$. Denote K as the number of clusters, and $A_1, \dots, A_K$ as the index sets of the K disjoint clusters. The NCut objective function is defined as

$$\text{NCut}(A_1, \dots, A_K) = \sum_{k=1}^K \frac{\text{cutvol}(A_k; \mathbf{U})}{\text{cut}(A_k, A_k^c; \mathbf{U})}, \quad (1)$$

where A_k^c is the complement of A_k ,

$$\text{cutvol}(A_k; \mathbf{U}) = \sum_{i, l \in A_k} u_{il}, \quad (2)$$

and

$$\text{cut}(A_k, A_k^c; \mathbf{U}) = \sum_{i \in A_k, l \in A_k^c} u_{il}. \quad (3)$$

The NCut technique has a very intuitive format. It maximizes the within-cluster similarity $\text{cutvol}(\cdot)$ and minimizes the across-cluster similarity $\text{cut}(\cdot)$ simultaneously. It makes no assumption on the similarity measure and underlying data distributions and models and can lead to more accurate and more robust clustering results. In addition, the algorithms for NCut usually have solid mathematical properties that are supported by the graph theory, such as the eigendecomposition of Laplacian matrices. For example, they have the advantage of the possibility of defining upper or lower bounds for the objective function. The popular NCut algorithms include recursive two-way spectral clustering algorithms, direct k -way spectral clustering algorithms, and others.³⁰ They have been widely applied in image segmentation, document clustering, genetics, and many other fields, demonstrating significant computational and numerical advantages.

2.2 | Clustering analysis with assisted weighted NCut

Motivated by the recent successes of assisted analysis under other contexts, our goal is to conduct more accurate clustering using GE data with the assistance of regulator information. Two *weighted* adjacency matrices $U_{(w)} = (u_{(w)il})_{n \times n}$ and $V_{(w)} = (v_{(w)il})_{n \times n}$ are constructed for the n subjects using the p GEs and q CNAs, respectively. Different from the traditional adjacency matrix for NCut, which treats all variables equally, our proposed definition is based on the weighted Euclidean distance. Specifically,

$$u_{(w)il} = \frac{1}{\sqrt{\sum_{j=1}^p w_j^{(x)} (x_{ij} - x_{lj})^2}}, v_{(w)il} = \frac{1}{\sqrt{\sum_{j=1}^q w_j^{(z)} (z_{ij} - z_{lj})^2}},$$

where $\mathbf{w}^{(x)} = (w_1^{(x)}, \dots, w_p^{(x)})$ and $\mathbf{w}^{(z)} = (w_1^{(z)}, \dots, w_q^{(z)})$ are the unknown nonnegative weight vectors for GEs and CNAs, respectively, with $\|\mathbf{w}^{(x)}\|_2 = 1$ and $\|\mathbf{w}^{(z)}\|_2 = 1$. Here, we adopt the Euclidean distance because of its simplicity and popularity. Weighted adjacency matrices based on other distances can be defined in a similar manner.

Using similar notations as in the last section, we propose the AWNCut (assisted weighted NCut) objective function as

$$\begin{aligned} \text{AWNCut}(A_1, \dots, A_K) = & \sum_{k=1}^K \left[\frac{\text{cutvol}(A_k; U_{(w)})}{\text{cut}(A_k, A_k^c; U_{(w)})} + \tau \frac{\text{cutvol}(A_k; V_{(w)})}{\text{cut}(A_k, A_k^c; V_{(w)})} \right] \\ & + \frac{\lambda}{p+q} \left(\sum_{j=1}^p w_j^{(x)} \overline{\text{cor}}(X_{A_k, j}, Z_{A_k, \cdot}) + \sum_{j=1}^q w_j^{(z)} \overline{\text{cor}}(Z_{A_k, j}, X_{A_k, \cdot}) \right). \end{aligned}$$

Here

$$\overline{\text{cor}}(X_{A_k, j}, Z_{A_k, \cdot}) = \frac{1}{q} \sum_{j'=1}^q \left| \text{cor}(X_{A_k, j}, Z_{A_k, j'}) \right|, \quad (4)$$

$$\overline{\text{cor}}(Z_{A_k, j}, X_{A_k, \cdot}) = \frac{1}{p} \sum_{j'=1}^p \left| \text{cor}(Z_{A_k, j}, X_{A_k, j'}) \right|, \quad (5)$$

and $\text{cor}(\mathbf{c}, \mathbf{d})$ is the Pearson correlation coefficient of two vectors $\mathbf{c}$ and $\mathbf{d}$. τ and $\lambda \geq 0$ are two data-dependent tuning parameters. With fixed tunings and K , we propose maximizing the AWNCut objective function, subject to the constraint

$$\mathbf{w}^{(x)} \geq 0, \mathbf{w}^{(z)} \geq 0, \|\mathbf{w}^{(x)}\|_2 = 1, \|\mathbf{w}^{(z)}\|_2 = 1,$$

which will lead to the cluster membership $A_1, \dots, A_K$ and weights $\mathbf{w}^{(x)}$ and $\mathbf{w}^{(z)}$.

Rationale.—The proposed method has been motivated by the following considerations. The first two terms are similar to that under a standard NCut, with one using GEs and the other using regulators. In the literature,³¹ regulators have been used to cluster samples. As such, the second term is sensible. With the regulation relationships between GEs and regulators, here, we reinforce that they generate the same clustering structure. We do note that information in GEs and regulators is not necessarily identical. However, as they usually contain significantly overlapping information, reinforcing identical clusterings, although may be slightly stringent, can realize effective information borrowing and improve clustering. Loosely speaking, the proposed method searches for clusters with distinct genetic profiles in a sense possibly broader than GEs. This strategy of reinforcing “concordance” between GE- and regulator-based analysis has also been proposed in the published assisted analysis¹⁸ and shown to be effective. Note that, with a significantly different analysis goal, the structure of our objective function and how concordance is reinforced differ significantly from the existing studies.¹⁸ Significantly, different from the standard NCut, weighted distances are calculated with data-dependent weights. The intuition is that the weights can effectively downweigh GEs and regulators irrelevant for clustering. The idea of weighting has also been proposed in the literature, for example, in sparse Kmeans⁶ where a Lasso penalty is imposed for identifying important variables. Different from the sparse Kmeans and other sparse methods, our resulted weights are dense, allowing for a large number of GEs (and regulators) to have small but nonzero contributions. The ℓ_2 -norm constraint also makes computation more feasible than the ℓ_1 constraint. If sparsity is desired, a simple thresholding can be applied to the weights. Another significant advancement of the proposed method is the last two terms, which promote the set of important GEs to be correlated with the set of important regulators. This is sensible considering the GE-regulator regulations. We note that there are other, possibly more informative, ways of describing the GE-regulator interconnections than correlation. However, correlation has been adopted in a large number of studies³² and shown to be effective. In addition, we have experimented with a few other options and found the proposed one to be computationally the most feasible.

In a practical study, there is no guarantee that all relevant regulators are measured. Tunings τ and λ are introduced to accommodate different “degrees of relevance” of regulators. More specifically, τ is used to adjust the relative importance of CNAs. It can downweigh the contribution of regulators in the case of not highly informative. The tuning parameter λ is used to balance the “cut” (the first and second terms) and regulation relationships. In the worst scenario where $X \perp Z$ (that is, the measured set of regulators is not relevant for the GEs), then with $\tau, \lambda \sim 0$, the proposed approach goes back to that using GEs only. Intuitively, this ensures that performance of the proposed approach will be no worse than that of NCut. It is noted that different values of τ may be adopted for different clusters to accommodate the differences across clusters. However, this may significantly increase computational cost and reduce stability. In our study, we use the same τ for all clusters, which generates satisfactory numerical results. The value of λ is divided by $p + q$ to achieve comparability of the two parameters. It is noted that the division is not critical and does not influence the final clustering results.

2.3 | Computation

We propose using the simulated annealing (SA) technique, which is iterative, to optimize the AWNCut objective function. Denote b as the index of iteration and B as the maximum iteration times. At iteration b , let $\{A_1^{(b)}, \dots, A_K^{(b)}\}$ be the clustering result and $\text{AWNCut}(b)$ be the value of the objective function. Define the temperature function as $T(b) = L \log(b + 1)$. Following published literature,³³ we set $L = 1000$. The proposed algorithm is presented in Algorithm 1.

Algorithm 1

Initialize $(w^{(x)})^{(0)}$ and $(w^{(z)})^{(0)}$ as $(w_1^{(x)})^{(0)} = (w_2^{(x)})^{(0)} = \dots = (w_p^{(x)})^{(0)} = \frac{1}{\sqrt{p}}$ and $(w_1^{(z)})^{(0)} = (w_2^{(z)})^{(0)} = \dots = (w_q^{(z)})^{(0)} = \frac{1}{\sqrt{q}}$. That is, the “prior” is noninformative. Randomly initialize $A_1^{(0)}, \dots, A_K^{(0)}$. Set $b = 0$, $pace_b^{(x)} = \frac{1}{10(1 + [b/10])\sqrt{p}}$ and $pace_b^{(z)} = \frac{1}{10(1 + [b/10])\sqrt{q}}$.

for $b = 1; b < B; b++$ **do**

With $w^{(x)}$ and $w^{(z)}$ fixed at $(w^{(x)})^{(b-1)}$ and $(w^{(z)})^{(b-1)}$, optimize with respect to the cluster memberships:

Use $(w^{(x)})^{(b-1)}$ and $(w^{(z)})^{(b-1)}$ to calculate matrices $U_{(w)}^{(b)}$ and $V_{(w)}^{(b)}$, respectively.

for $k = 1; k < K; k++$ **do**

compute $p_k = \frac{\text{cutvol}(A_k; U_{(w)}^{(b)})}{\text{cut}(A_k, A_k^c; U_{(w)}^{(b)})} + \tau \frac{\text{cutvol}(A_k; V_{(w)}^{(b)})}{\text{cut}(A_k, A_k^c; V_{(w)}^{(b)})}$.

end for

Let $k(-) = \max_{k=1, \dots, K} p_k$ and $k(+) = \min_{k=1, \dots, K} p_k$.

For $A_{k(-)}^{(b)}$, compute the weighted Euclidean distance for each subject pair using GEs and CNAs, respectively. Select the pair with the largest weighted Euclidean distance for GEs and CNAs, respectively. Draw i randomly from the subjects in these two pairs. For $j \neq k(+), k(-)$, $A_j^{(b)} := A_j^{(b-1)}$, set

$$A_{k(+)}^{(b)} = A_{k(+)}^{(b-1)} \cup \{i\}, A_{k(-)}^{(b)} = A_{k(-)}^{(b-1)} \setminus \{i\}. \quad (6)$$

With $A_1, \dots, A_K$ fixed at $A_1^{(b)}, \dots, A_K^{(b)}$, optimize with respect to $w^{(x)}$ and $w^{(z)}$:

for $j = 1; j < p, j++$ **do**

compute $\sum_{k=1}^K \overline{\text{cor}} \left(X_{A_k^{(b)}j}, Z_{A_k^{(b)}} \right)$ as defined in (4)

end for

With the p computed values, separate the GEs into two groups $F_1^{(x)}$ and $F_2^{(x)}$ using an unsupervised approach (for example Kmeans). Without loss of generality, Assume that $F_1^{(x)}$ has a larger value of centroid. Conduct a similar procedure on the regulators to get $F_1^{(z)}$ and $F_2^{(z)}$.

Update

$$\left(\mathbf{w}_{F_1^{(x)}}^{(x)}\right)^{(b)} = \left(\mathbf{w}_{F_1^{(x)}}^{(x)}\right)^{(b-1)} + pace_b \left(\mathbf{w}_{F_2^{(x)}}^{(x)}\right)^{(b)} = \max\left(\left(\mathbf{w}_{F_2^{(x)}}^{(x)}\right)^{(b-1)} - pace_b, 0\right), \quad (7)$$

and

$$\left(\mathbf{w}_{F_1^{(z)}}^{(z)}\right)^{(b)} = \left(\mathbf{w}_{F_1^{(z)}}^{(z)}\right)^{(b-1)} + pace_b \left(\mathbf{w}_{F_2^{(z)}}^{(z)}\right)^{(b)} = \max\left(\left(\mathbf{w}_{F_2^{(z)}}^{(z)}\right)^{(b-1)} - pace_b, 0\right). \quad (8)$$

Normalize $\left(\mathbf{w}^{(x)}\right)^{(b)}$ and $\left(\mathbf{w}^{(z)}\right)^{(b)}$ to satisfy $\left\|\left(\mathbf{w}^{(x)}\right)^{(b)}\right\|_2 = 1$ and $\left\|\left(\mathbf{w}^{(z)}\right)^{(b-1)}\right\|_2 = 1$.

if AWNCut (b) - AWNCut ($b - 1$) **then**

keep $A^{(b)}$ as it is;

else keep $A^{(b)}$ as it is with probability $\exp\left(-\frac{\text{AWNCut}(b) - \text{AWNCut}(b-1)}{T(b)}\right)$, and otherwise $A^{(b)} = A^{(b-1)}$.

end if

if $\frac{\sum_{j=1}^{p+q} |w_j^{(b)} - w_j^{(b-1)}|}{\sum_{j=1}^{p+q} |w_j^{(b-1)}|} < 10^{-4}$ or $b = B$, where $\mathbf{w} = (w_1^{(x)}, \dots, w_p^{(x)}, w_1^{(z)}, \dots, w_q^{(z)})$, **then**

End Loop

end if

end for

The SA technique has been widely adopted in the literature (particularly, including the ANCut study) and shown to have satisfactory performance. Our objective function is considerably more complicated and involves the cluster memberships and weights. We iteratively optimize over them. When updating the cluster memberships (6), we encourage the clusters to have similar sizes, where a new member is deleted from a cluster with a larger “cut” and added to a cluster with a smaller “cut”. It is noted that this strategy may be somewhat subjective and can be adjusted if necessary. In (7) and (8), the weights for GEs and CNAs are updated according to their degrees of regulation relationships. Kmeans is used to separate GEs (CNAs) into two groups, where we note that the adopted Kmeans is not essential and other clustering approaches can also be applied. The SA algorithm is a probabilistic method based on the discrete time inhomogeneous Markov chain. It has been shown to converge under weak assumptions, which is supported by the convergence theory of the irreducible and aperiodic Markov chain.³⁴ The speed of convergence has been also extensively discussed in the literature³⁴ and expected to be satisfactorily fast. Convergence is achieved in all of our numerical examples.

Tuning parameter selection.—As with other clustering methods, the proposed method involves the number of clusters K . In addition, it also involves τ and λ , which balance the weighted NCut objective function for GEs and information borrowing from regulators. We

select τ and λ using the Davies-Bouldin index (DBI).³⁵ With cluster memberships $A_1, \dots, A_K$ and weights $w^{(x)}$ and $w^{(z)}$, let

$$M_{kk'} = \sum_{i \in A_k, l \in A_{k'}} \sqrt{\sum_{j=1}^p w_j^{(x)} (x_{ij} - x_{lj})^2 + \sum_{j=1}^q w_j^{(z)} (z_{ij} - z_{lj})^2}$$

and

$$S_k = \sum_{i, l \in A_k} \sqrt{\sum_{j=1}^p w_j^{(x)} (x_{ij} - x_{lj})^2 + \sum_{j=1}^q w_j^{(z)} (z_{ij} - z_{lj})^2}$$

be the across- and within-cluster dissimilarity based on the Euclidean distance, respectively. Then, the DBI is defined as

$$\text{DBI} = \frac{1}{K} \sum_{k=1}^K \max_{k' \neq k} \left(\frac{M_{kk'}}{S_k + S_{k'}} \right), \quad (9)$$

with a larger value indicating a better clustering performance. For choosing K , we adopt the GAP statistic,³⁶ which has been extensively applied in the literature, shown to have satisfactory performance and can be realized using the R function *clusGap*. To determine the optimal values of all tunings, we conduct a grid search.

Computation realization.—To facilitate data analysis, we have developed R code implementing the proposed method and made it publicly available at www.github.com/shuanggema, where users can also find an easy-to-understand manual and demonstrating examples.

2.4 | A closer look

To better appreciate the operating characteristics of AWNCut, we simulate one dataset under simulation Scenario S1 with $a = 1$. Here, the first 20 GEs and CNAs are informative for defining the sample clusters. Details on the data settings are described in the next section. Beyond the proposed method, we also consider the sparse Kmeans (SKmeans) approach. Kmeans is one of the most popular clustering methods, and we adopt its sparse version, which is expected to generate better results under this particular simulation setting. For AWNCut, a grid search generates the optimal tuning values $(\tau, \lambda, K) = (0.15, 3, 3)$. As a four-dimensional plot is hard to generate and visualize, in Figures A1 and A2 in the Appendix, for each tuning, we fix the other tunings at their optimal value and plot the tuning parameter selection statistic. The optimal points are clearly identified. In Figure 1, we plot the estimated weights under the two methods. Under the proposed methods, the irrelevant GEs and CNAs have weights very close to zero. A simple thresholding can easily separate the relevant GEs/CNAs from the irrelevant. The separation of weights is also observed for SKmeans, however, not as distinct as the proposed method. This better identification of

relevant GEs/CNAs leads to a better clustering. For this particular dataset, AWNCut and SKmeans have clustering errors $M_{\text{error}} = 0.271$ and 0.439 , respectively. More definitive results are presented in the next section.

3 | SIMULATION

Simulation is conducted to evaluate performance of the proposed method and compare with competing alternatives. Specifically, the following most relevant alternatives are considered. **NCutXZ**, which is the same as the proposed, except that all weights are set as equal. That is, there is no mechanism for discriminating variables. Comparing with this approach can reveal the value of adding weights. **WNCut**, which is a weighted NCut approach based on the stacked data $(\mathbf{X}, \mathbf{Z})$. Comparing with this approach can reveal the value of assisted analysis. **NCut**, **Kmeans**, **SKmeans**, and **Spectral**, which analyze the stacked data $(\mathbf{X}, \mathbf{Z})$ using NCut, standard Kmeans, SKmeans, and spectral clustering,³⁷ respectively. These “standard” approaches have been extensively used in the literature and demonstrated to have satisfactory results. We acknowledge that there are (many) other methods that can be applied to the simulated data. The above alternatives have frameworks closest to the proposed/are the most popular, and comparing with them can more directly establish the merit of the proposed assisted and weighted analysis.

When assessing performance, the following measure is adopted. Given a clustering result $A_1, \dots, A_K$, we construct a matrix $\mathbf{A} = (a_{ij})_{n \times n}$, where the (i, j) th element $a_{ij} = 1$ if $i, j \in A_k$, $k = 1, \dots, K$ and 0 otherwise. Let $\mathbf{A}_T$ and $\hat{\mathbf{A}}$ be the corresponding matrices for the true and estimated clusters, respectively. Then, the error rate of the estimated clustering result is defined as the diversity between $\mathbf{A}_T$ and $\hat{\mathbf{A}}$,

$$M_{\text{error}} = \frac{\sum_{i,l} (\mathbf{A}_T \odot \hat{\mathbf{A}})_{il}}{\sum_{i,l} (\mathbf{A}_T)_{il}}, \quad (10)$$

where $\odot$ is the component wise product, and a smaller value of M_{error} suggests a better clustering result.

We consider the following simulation settings. (a) $K = 3$, $p = 500$, and $q = 300$. (b) Both GEs and CNAs have p_0 important components. (c) For the k th cluster, the important CNAs are generated from a transformed multivariate normal distribution $\text{sign}(\mathbf{G})\exp(\mathbf{G})$, where $\mathbf{G}$ follows a multivariate normal distribution with marginal means μ_k and covariance matrix Σ . Here, Σ has diagonal elements equal to 1 and off-diagonal elements equal to 0.8. The important GEs are regulated by the important CNAs through a linear regression model $\mathbf{X} = \beta\mathbf{Z} + \epsilon$, where each element of ϵ is iid from a standard normal distribution, and β is the coefficient matrix quantifying the strengths of regulation. (d) The noisy GEs and CNAs are generated in a similar manner as for the important ones. Important and noisy GEs/CNAs are independent. The above settings have been motivated by the following considerations. In TCGA and other studies, processed CNA data are analyzed, which have continuous distributions. We generate from transformed normal distributions, which are unimodal but

not symmetric and may better mimic real data. Different clusters have mean shifts in the untransformed variables. Important CANs are closely related to each other and regulate important GEs through a linear regression model, which has been extensively adopted in the literature. Under the assumption that the noisy GEs/CNAs have functionalities not related to clustering, we generate them independent of the important ones. This type of independence has also been extensively considered in the literature.

The following scenarios are considered with different β , p_0 , μ_k , and n_k . Among them, β and p_0 describe the regulation relationships. μ_k describes the level of separation between clusters. In clustering analysis, balancing between clusters is of interest, has important implications for some methods, and is also examined here.

- S1,** which has each element of β iid from Uniform $[-0.5, 0.5]$, $p_0 = 20$, $n_1 = n_2 = n_3 = 30$, and $\mu_1 = -a$, $\mu_2 = 0$, and $\mu_3 = a$ with various a values. Here, our goal is to evaluate how performance of the proposed method depends on the separation of clusters.
- S2,** which is the same as Scenario S1, except that the sizes of the three clusters are $n_1 = 3$, $n_2 = 6$, and $n_3 = 81$. This highly unbalanced setting examines the dependence of the proposed method on balancing.
- S3,** which is the same as Scenario S1, except with various p_0 values. This scenario examines dependence of the proposed method on the signal-to-noise ratios of GEs and CNAs.
- S4,** which has $p_0 = 20$, $n_1 = n_2 = n_3 = 30$, $\mu_1 = -0.5$, $\mu_2 = 0$, and $\mu_3 = 0.5$, and two different settings of β . Under both settings, β has t_0 elements equal to zero with various t_0 values. In addition, each nonzero element of β is iid from Uniform $[-0.5, 0.5]$ and Uniform $[-0.05, 0.05]$ under the first (S4.1) and second (S4.2) settings, respectively. This scenario examines whether performance of the proposed method depends on the strengths of regulation.
- S5,** which is the same as Scenario S1, except with various values of n . This scenario examines dependence of the proposed method on the sample size.

There are a total of 39 simulation settings. For each setting, 500 replicates are simulated, and summary statistics are computed. Results for Scenarios S1 and S2 are shown in Table 1. The rest of the results are shown in the Appendix. It is observed that across all simulation settings, the proposed method has competitive performance. Under Scenario S1, the proposed method can more accurately identify clusters with various signal levels. For example, with signal level $a = 1.0$, the proposed method has $M_{\text{error}} = 0.257$, compared to 0.271 (NCutXZ), 0.324 (WNCut), 0.320 (NCut), 0.393 (Kmeans), 0.439 (SKmeans), and 0.322 (Spectral). The proposed method behaves better compared to NCutXZ, which provides a strong support to the proposed weighted strategy. It also has better performance than WNCut, which suggests that the assistance of CNAs can improve sample clustering. Compared to the proposed method, NCut, Kmeans, SKmeans, and Spectral, which analyze the stacked data (X, Z) directly, have inferior performance. This suggests that it pays off to quantify the interconnections among GEs and regulators. Under the highly unbalanced Scenario S2, all methods perform worse, which is as expected. Similar to Scenario S1, the

proposed method is observed to have superior performance. The advantage of the proposed method gets more prominent with an increase in the signal-to-noise ratios of GEs and CNAs. For example, under Scenario S3 (Table A1) with $a = 0.3$ and $p_0 = 40$, the M_{error} values are 0.266 (proposed), 0.290 (NCutXZ), 0.328 (WNCut), 0.326 (NCut), 0.438 (Kmeans), 0.442 (SKmeans), and 0.362 (Spectral). Under Scenario S4 (Table A2), performance of the proposed method is observed to decay with the increase of t_0 and decrease of the nonzero value of β . This is reasonable as the effectiveness of the proposed assisted analysis is less justified with the weakening of the regulation relationship. Under Scenario S5 (Table A3) with various values of sample size, it is observed that the proposed method still demonstrates superior or comparable performance compared to the six alternatives. For example, with $n = 150$ and $a = 1$, the M_{error} values are 0.274 (proposed), 0.289 (NCutXZ), 0.326 (WNCut), 0.328 (NCut), 0.422 (Kmeans), 0.410 (SKmeans), and 0.281 (Spectral). We have also experimented with a few other simulations settings, and the observed patterns are similar (details omitted).

Computational cost.

The proposed computational algorithm only involves simple updates, making the proposed approach computationally affordable. With fixed K , Table A4 (Appendix) provides the average computational time for the analysis of one replicate under Scenario S1 with $a = 1$. We examine the results with various values of n , p , and q . For both AWNCut and NCutXZ, we consider seven and five values of λ and τ , respectively. Due to the two-dimensional grid search for the optimal λ and τ and the update of the weights, the proposed method is relatively more time consuming than the alternatives. However, it is still computationally feasible and can accommodate data with high dimensionality. For a simulated dataset with $n = 150$, $p = 600$, and $q = 400$, the proposed analysis takes about 2 minutes on a laptop with standard configurations, with all the tuning parameters fixed. The process of tuning parameters via grid search can be realized in a parallel manner, dramatically reducing computational time.

4 | DATA ANALYSIS

We analyze TCGA data on cutaneous melanoma (SKCM) and lung adenocarcinoma (LUAD) and search for groups of subjects with distinct GE patterns with the assistance of CNA measurements. The processed level 3 data are analyzed, which are downloaded from TCGA Provisional using the R package *cgdscr*. For GEs, to be analyzed are the Z-scores, which are collected using the IlluminaHiSeq RNAseq V2 platform and have been lowess-normalized, log-transformed, and median-centered. They represent the GE status (up or down regulated) in tumor samples relative to normal tissues. The CNA measurements are collected using the Affymetrix Genome-wide Human SNP array 6.0 platform. The loss and gain levels of copy number changes of tumors compared to normal tissues are identified using segmentation analysis and expressed in the log2 transformed form.

4.1 | SKCM data

For SKCM data, preprocessing similar to that in the literature²⁴ is conducted, which includes subject matching and missing data imputation. In addition, although the proposed method is

potentially applicable to a large number of GEs, to improve stability, a marginal screening is conducted, selecting 1000 GE and CNA measurements. Furthermore, 382 GEs and CNAs are identified as having well-defined GO (Gene Ontology) terms using GOTerm Finder.³⁸ This screening is also strongly motivated by the consideration that the number of relevant genes is small, which is coherent with our weighted strategy. The number of subjects is 367.

The GAP approach identifies eight clusters for the proposed and alternative methods. The sizes of the clusters are presented in Table 2. More detailed information is available from the authors. To compare across methods, we try to match the clusters identified using different methods via correlation. Table 2 suggests that different methods lead to different clustering structures. It is noted that the Kmeans and SKmeans methods generate much unbalanced clusterings, while the proposed method generates more balanced. In practical data analysis, it is hard to objectively evaluate the accuracy of clustering without the true underlying structures. To provide partial support, we compute the values of DBI using a resampling-based method. First, the subjects are randomly partitioned into two disjoint subsets with equal size. Next, for each method, the optimal tuning parameters are selected on the first subset, with which we then compute the DBI on the

Advancing from some methods, the proposed can generate weights for GEs and regulators. In Figure 2, we plot the estimated weights. It is observed that there are GEs and regulators clearly with larger weights. That is, they can be discriminated into classes with different importance. We search the GO terms for the GEs with larger weights and find that the enriched terms include *regulation of biological process*, *regulation of cellular process*, *regulation of transportation*, *immune response*, *defense response*, and others, all of which have been implicated to have important roles in melanoma and other cancers. This biological soundness also provides support to the proposed method. As with other clustering methods, the proposed separates subjects into groups with different GE patterns. As an indirect evaluation, we examine whether such differences in GEs are associated with important clinical outcomes/phenotypes. In Figure A3 in the Appendix, we plot the estimated densities of log(Breslow thickness) for the eight clusters, where Breslow thickness is a widely examined outcome variable and has important implications for melanoma prognosis. We also present the Kaplan-Meier curves of survival time and distributions of the mutation subtypes for the eight clusters in Figures A4 and A5. The four mutation subtypes are referred to as mutant BRAF, mutant RAS, mutant NF1, and triple wild type (TWT), which are well established in the literature.³⁹ The mutation data are only available for 244 subjects. For all the three analyses, some differences across clusters are observed. It is noted that the differences are not “dramatic”, which is sensible, as the proposed method conducts unsupervised clustering not tailored to these outcome variables.

4.2 | LUAD data

We conduct a similar prescreening, and 373 GEs and CNAs are selected for downstream analysis. Data are available on 503 subjects. With the proposed method, three clusters are generated using the GAP statistic. Table 2 provides the sizes of the clusters for the proposed method and six alternatives with $K = 3$. It again suggests that AWNCut generates results different from using the alternatives. With the GAP approach, the six alternatives generate

two (NCutXZ), seven (WNCut), three (NCut), eight (Kmeans), three (SKmeans), and seven (Spectral) clusters, respectively. The DBI values are computed using the same approach as described above. The proposed method has the average DBI 2.099, compared to 2.086 (NCutXZ), 1.041 (WNCut), 1.043 (NCut), 1.237 (Kmeans), 1.300 (SKmeans), and 1.187 (Spectral). The improvement in DBI suggests the effectiveness of the proposed method.

We present the estimated weights of GEs and CNAs in Figure 2. It is observed that the variables naturally form two groups, where the one with much larger weights has more contribution to clustering. The GO term enrichment analysis finds that some terms are significantly over-represented among the GEs with larger weights, including *regulation of mRNA metabolic process*, *mRNA processing*, *pyridine nucleotide metabolic process*, *nucleotide-excision repair*, *positive regulation of telomere maintenance*, and others, which may provide certain biological support to the clustering. We further examine the differences in survival time and disease subtypes among subjects in different clusters in Figures A6 and A7. The subtypes include Proximal proliferative (PP), Proximal inflammatory (PI), and Terminal respiratory unit (TRU), which are associated with KRAS mutation, co-mutation of NF1 and TP53, and EGFR mutation, respectively.⁴⁰ The mutation data are available for 129 subjects. Some differences similar to those observed in the previous analysis are observed.

4.3 | Stability analysis based on TCGA data

We further conduct simulation based on the TCGA data to examine the stability of the proposed method, in terms of clustering detection and variable selection. Specifically, two scenarios are considered. In the first scenario (D1), we add 10% p and 10% q variables into the GEs and CNAs, respectively, where the variables are independently generated from $N(0, 1)$. In the second scenario (D2), 10% of the subjects are randomly removed from the datasets. Then, we evaluate the stability of clustering detection using

$$M_{\text{stability}} = \sum_{i,l} |\hat{A}^{(s)} - \hat{A}^{(0)}|_{il}^2 / n^2,$$

following the definition proposed in the literature,²⁴ where $\hat{A}^{(s)}$ and $\hat{A}^{(0)}$ are the corresponding matrices for the estimated clusters on the simulated and original datasets. A smaller $M_{\text{stability}}$ indicates higher stability. The mean and standard deviation of $M_{\text{stability}}$ over 200 randomly replicates are provided in Table A5. Although with more tuning parameters, the proposed method is observed to still have satisfactory clustering stability. For AWNCut, WNCut, and SKmeans, which have the variable selection component, we also examine the selection stability by calculating the observed occurrence index (OOI),⁴¹ which measures the identified probability of each of the top 60 variables with the largest weights. The summary results are presented in Table A5. It is observed that the proposed method has much higher selection stability. For example, for Scenario D1 based on the SKCM dataset, the proposed method has average OOI = 0.997, while WNCut and SKmeans only have 0.069 and 0.145, respectively. This stability analysis provides a certain degree of confidence to the numerical results of the proposed method.

5 | DISCUSSION

Clustering is a fundamental analysis in the research and treatment of complex diseases. In this study, we have developed a novel assisted clustering strategy. The approach of reinforcing certain “concordance” in the GE- and regulator-based analysis while accounting for the regulations shares a similar spirit with but also advances significantly from that in Chai et al¹⁸ and other existing assisted analyses. A closer examination of the proposed objective function suggests that the proposed strategy can potentially be coupled with other clustering measures and be broadly applicable. This study has also significantly extended the NCut technique, which has found more and more applications in genetic studies. Simulations and data analysis suggest competitive performance of the proposed analysis.

In published studies, some clustering approaches can be applied to detect gene clusters in the same manner as that for subject group identification. However, due to the unique characteristics of multidimensional data, the proposed subject clustering approach cannot be used to cluster GEs directly. The techniques for accommodating regulation relationships between GEs and regulators may differ significantly between subject clustering and gene clustering. For instance, in the recent ANCut approach,²⁴ the goal is to conduct more effective clustering of GEs with the assistance of information from regulators. For this purpose, the GEs are first decomposed into two components using penalized regression. The first one contains information regulated by CNAs, and the second one contains information regulated by the other regulators and those affected by other mechanisms besides regulation. Then, the minimization of NCut measures for both components of GEs is developed. With a significantly different goal, the proposed method takes advantage of the regulation relationships between GEs and regulators to achieve more accurate identification of subject groups. The GEs and their regulators are reinforced to generate the same clustering structure, leading to clusters with distinct genetic profiles broader than GEs. In addition, the correlations between GEs and CNAs are incorporated into the objective function to accommodate the GE-regulator regulations, different from the decomposition framework. Recently, for identifying subject groups, the sparse clustering framework has attracted much attention.^{6,7,42} On one hand, with high-dimensional omics data, differences across subject clusters are usually only associated with to a small fraction of variables. The true clustering structures may be missed if too many noninformative variables are used in clustering. On the other hand, the “large $p(q)$, small n ” paradigm may make some traditional clustering methods infeasible. For example, the $p \times p$ covariance matrix cannot be readily estimated with fewer subjects in the model-based clustering with the Gaussian mixture model.¹¹ As such, we have developed the assisted clustering under the sparse framework to detect the true cluster structures more effectively.

This study can be potentially extended in multiple directions. We have assumed that GE and regulator measurements are available for all subjects. The proposed method can be potentially extended to accommodate the “partially matched” scenario, where some subjects only have GE measurements. As discussed above, it can be potentially interesting to couple the proposed assisted strategy with other clustering techniques (Kmeans, etc), which may demand a separate development. There are other, possibly more effective, ways to describe the GE-regulator relationship. The proposed has been found to be computationally feasible

and led to satisfactory results. The tuning parameters are selected using a grid search on the predefined three-dimensional parameter domain. There may be theoretical ranges on τ and λ , similar to those reported in other studies, such as the penalized regression.⁴³ In this study, we have mostly focused on methodological development and implementation. Theoretical investigation is postponed to future study. The determination of the optimal number of clusters K is a major challenge in cluster analysis and has attracted many concerns. There is little research on the selection criterion of the optimal number of clusters for multidimensional data clustering analysis. In our study, we adopt the GAP approach that has been the most common choice in single data set-based clustering analysis. It has been also suggested in recent clustering studies with multidimensional data and shown to behave well in numerical analyses.^{44,45} Thus, it may be reasonable to conjecture that the GAP statistic is also valid for multidimensional data. Detailed studies are deferred to future investigation. In data analysis, the proposed method leads to a larger DBI measure, and the GEs with larger weights are biologically sensible. A more direct confirmation would involve examining the pathological features of the identified clusters and quantifying their differences. However, this would demand additional studies and is beyond this article.

ACKNOWLEDGEMENTS

We thank the associate editor and three reviewers for their careful review and insightful comments, which have led to a significant improvement of this article. This work was partly supported by the National Institutes of Health (CA216017 and CA204120), by the National Natural Science Foundation of China (71771211, 61402276, and 91546202), and by the Bureau of Statistics of China (2016LD01).

Funding information

National Institutes of Health, Grant/Award Number: CA216017 and CA204120; National Natural Science Foundation of China, Grant/Award Number: 71771211, 61402276, and 91546202; National Bureau of Statistics of China, Grant/Award Number: 2016LD01

APPENDIX

ADDITIONAL NUMERICAL RESULTS

FIGURE A1.

Simulation: Davies-Bouldin index (DBI) as a function of λ and τ

FIGURE A2.

Simulation: GAP statistic as a function of K [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE A3.

Analysis of SKCM data: densities of log(Breslow thickness) for the eight clusters [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE A4.

Analysis of SKCM data: Kaplan-Meier curves of survival time for the eight clusters [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE A5.

Analysis of SKCM data: distributions of the mutation subtypes (BRAF, RAS, NF1, and TWT) for the eight clusters [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE A6.

Analysis of LUAD data: Kaplan-Meier curves of survival time for the three clusters [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE A7.

Analysis of LUAD data: distributions of the mutation subtypes (PP, PI, and TRU) for the three clusters [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE A1

Simulation scenario S3. In each cell, mean (sd) based on 500 replicates

a	P_0	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
0.3	3	0.281(0.021)	0.295(0.020)	0.329(0.005)	0.330(0.005)	0.421(0.076)	0.445(0.082)	0.353(0.041)
	20	0.272(0.032)	0.289(0.021)	0.328(0.010)	0.328(0.010)	0.419(0.062)	0.447(0.064)	0.359(0.035)
	40	0.266(0.036)	0.290(0.020)	0.328(0.008)	0.326(0.008)	0.438(0.059)	0.442(0.074)	0.362(0.037)
0.5	3	0.277(0.024)	0.297(0.018)	0.329(0.007)	0.328(0.006)	0.428(0.096)	0.431(0.050)	0.351(0.038)
	20	0.270(0.032)	0.286(0.024)	0.327(0.007)	0.326(0.008)	0.428(0.065)	0.450(0.088)	0.356(0.043)

a	P_0	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
1.0	40	0.265(0.033)	0.287(0.021)	0.326(0.008)	0.325(0.009)	0.425(0.064)	0.444(0.089)	0.353(0.039)
	3	0.274(0.025)	0.292(0.023)	0.330(0.006)	0.329(0.006)	0.429(0.095)	0.477(0.142)	0.320(0.041)
	20	0.261(0.037)	0.279(0.025)	0.324(0.009)	0.320(0.012)	0.439(0.118)	0.439(0.108)	0.322(0.043)
	40	0.251(0.039)	0.274(0.025)	0.317(0.015)	0.317(0.015)	0.443(0.134)	0.423(0.097)	0.323(0.045)

TABLE A2

Simulation scenario S4. In each cell, mean (sd) based on 500 replicates

	t_0	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
S4.1	0	0.268(0.022)	0.280(0.031)	0.321(0.008)	0.333(0.005)	0.463(0.071)	0.520(0.108)	0.325(0.052)
	5	0.269(0.017)	0.290(0.016)	0.326(0.009)	0.328(0.008)	0.473(0.060)	0.480(0.063)	0.353(0.063)
	10	0.272(0.015)	0.292(0.008)	0.327(0.004)	0.325(0.004)	0.448(0.041)	0.439(0.044)	0.341(0.018)
S4.2	0	0.270(0.032)	0.286(0.024)	0.327(0.007)	0.326(0.008)	0.428(0.065)	0.439(0.108)	0.356(0.040)
	5	0.276(0.022)	0.288(0.032)	0.325(0.005)	0.323(0.009)	0.402(0.049)	0.452(0.117)	0.306(0.029)
	10	0.281(0.026)	0.293(0.014)	0.323(0.007)	0.316(0.011)	0.440(0.058)	0.455(0.124)	0.325(0.027)

TABLE A3

Simulation scenario S5. In each cell, mean (sd) based on 500 replicates

a	n	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
0.5	60	0.282(0.028)	0.298(0.014)	0.329(0.008)	0.326(0.007)	0.439(0.051)	0.439(0.047)	0.354(0.038)
	120	0.280(0.030)	0.290(0.017)	0.326(0.005)	0.327(0.006)	0.462(0.042)	0.465(0.039)	0.339(0.027)
	150	0.295(0.014)	0.273(0.040)	0.329(0.006)	0.327(0.004)	0.390(0.032)	0.400(0.038)	0.343(0.032)
	210	0.263(0.016)	0.275(0.015)	0.328(0.002)	0.326(0.003)	0.455(0.027)	0.460(0.016)	0.320(0.024)
1.0	60	0.264(0.016)	0.263(0.021)	0.319(0.011)	0.314(0.012)	0.340(0.055)	0.319(0.091)	0.288(0.048)
	120	0.270(0.018)	0.282(0.018)	0.319(0.013)	0.304(0.017)	0.478(0.091)	0.500(0.168)	0.317(0.023)
	150	0.274(0.016)	0.289(0.013)	0.326(0.005)	0.328(0.008)	0.422(0.052)	0.410(0.072)	0.281(0.026)
	210	0.260(0.024)	0.262(0.018)	0.318(0.010)	0.319(0.005)	0.600(0.201)	0.503(0.182)	0.374(0.206)
1.5	60	0.246(0.034)	0.269(0.015)	0.309(0.007)	0.307(0.019)	0.364(0.095)	0.475(0.208)	0.292(0.057)
	120	0.252(0.041)	0.256(0.028)	0.324(0.010)	0.321(0.011)	0.525(0.192)	0.514(0.245)	0.298(0.054)
	150	0.262(0.028)	0.256(0.026)	0.325(0.008)	0.318(0.015)	0.482(0.092)	0.523(0.099)	0.318(0.038)
	210	0.264(0.021)	0.263(0.017)	0.317(0.010)	0.320(0.005)	0.607(0.208)	0.515(0.180)	0.362(0.204)

TABLE A4

Computational time (minutes) for scenario S1 with $a = 1$. In each cell, mean based on 500 replicates

p	q	n	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
500	300	90	36.338	4.898	0.554	0.260	0.0003	0.116	0.004
		120	39.669	4.991	0.926	0.447	0.0005	0.135	0.006
		150	50.414	6.140	1.435	0.711	0.0006	0.204	0.010

<i>p</i>	<i>q</i>	<i>n</i>	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
		90	61.493	7.814	0.684	0.318	0.0003	0.151	0.005
600	400	120	64.459	8.392	1.158	0.555	0.0005	0.201	0.008
		150	76.617	9.468	1.779	0.866	0.0008	0.254	0.010

TABLE A5

Stability analysis based on The Cancer Genome Atlas data. For $M_{\text{stability}}$, the mean (sd) based on 200 replicates are presented in each cell. For observed occurrence index, the mean (sd) for the top 60 variables are presented in each cell

Dataset	Scenario	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
$M_{\text{stability}}$								
SKCM	D1	0.216(0.001)	0.214(0.001)	0.216(0.001)	0.216(0.001)	0.302(0.019)	0.234(0.093)	0.182(0.004)
	D2	0.215(0.007)	0.214(0.001)	0.216(0.001)	0.215(0.001)	0.293(0.019)	0.212(0.127)	0.185(0.006)
LUAD	D1	0.334(0.002)	0.333(0.001)	0.332(0.001)	0.333(0.001)	0.607(0.122)	0.479(0.005)	0.351(0.003)
	D2	0.409(0.013)	0.391(0.013)	0.443(0.001)	0.444(0.001)	0.302(0.043)	0.580(0.009)	0.311(0.012)
OOI								
SKCM	D1	0.996(0.032)		0.069(0.022)			0.145(0.200)	
	D2	0.997(0.020)		0.073(0.083)			0.048(0.049)	
LUAD	D1	1.000(0.000)		0.072(0.027)			0.008(0.091)	
	D2	0.999(0.009)		0.081(0.029)			0.487(0.424)	

Abbreviation:

AWNCut   assisted weighed normalized cut

REFERENCES

- Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. *Bioinformatics*. 2016;32(12):1832–1839. [PubMed: 26873929]
- Marigorta UM, Denson LA, Hyams JS, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet*. 2017;49(10):1517–1521. [PubMed: 28805827]
- Sharabi-Schwager M, Or E, Ophir R. ctsGE—clustering subgroups of expression data. *Bioinformatics*. 2017;33(13):2053–2055. [PubMed: 28334165]
- Felipe De Sousa EM, Wang X, Jansen M, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med*. 2013;19(5):614–618. [PubMed: 23584090]
- Nidheesh N, Nazeer KAA, Ameer PM. An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data. *Comput Biol Med*. 2017;91:213–221. [PubMed: 29100115]
- Witten DM, Tibshirani R. A framework for feature selection in clustering. *J Am Stat Assoc*. 2010;105(490):713–726. [PubMed: 20811510]
- Sun W, Wang J, Fang Y. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron J Stat*. 2012;6:148–167.
- Liu H, Zhao R, Fang H, Cheng F, Fu Y, Liu Y-Y. Entropy-based consensus clustering for patient stratification. *Bioinformatics*. 2017;33(17):2691–2698. [PubMed: 28369256]

9. Chen GK, Chi EC, Ranola JMO, Lange K. Convex clustering: an attractive alternative to hierarchical clustering. *PLoS Comput Biol*. 2015;11(5). e1004228. [PubMed: 25965340]
10. Bouveyron C, Brunet-Saumard C. Model-based clustering of high-dimensional data: a review. *Comput Stat Data Anal*. 2014;71:52–78.
11. McNicholas PD, Murphy TB. Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*. 2010;26(21):2705–2712. [PubMed: 20802251]
12. Ahmad A, Fröhlich H. Towards clinically more relevant dissection of patient heterogeneity via survival-based Bayesian clustering. *Bioinformatics*. 2017;33(22):3558–3566. [PubMed: 28961917]
13. Yu Z, Wong H-S, Wang H. Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*. 2007;23(21):2888–2896. [PubMed: 17872912]
14. Yu Z, Luo P, You J, et al. Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Trans Knowl Data Eng*. 2016;28(3):701–714.
15. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. *Comput Biol Med*. 2008;38(3):283–293. [PubMed: 18061589]
16. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. *Nat Methods*. 2015;12(11):1033–1038. [PubMed: 26389570]
17. Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform*. 2014;16(2):291–303. [PubMed: 24632304]
18. Chai H, Shi X, Zhang Q, Zhao Q, Huang Y, Ma S. Analysis of cancer gene expression data with an assisted robust marker identification approach. *Genet Epidemiol*. 2017;41(8):779–789. [PubMed: 28913902]
19. Gross SM, Tibshirani R. Collaborative regression. *Biostatistics*. 2014;16(2):326–338. [PubMed: 25406332]
20. An X, Hu J, Do K-A. SIFORM: shared informative factor models for integration of multi-platform bioinformatic data. *Bioinformatics*. 2016;32(21):3279–3290. [PubMed: 27381342]
21. Luo C, Liu J, Dey DK, Chen K. Canonical variate regression. *Biostatistics*. 2016;17(3):468–483. [PubMed: 26861909]
22. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *Ann Appl Stat*. 2013;7(1):269–294. [PubMed: 24587839]
23. Hellton KH, Thoresen M. Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*. 2016;17(3):537–548. [PubMed: 26917056]
24. Hidalgo SJT, Wu M, Ma S. Assisted clustering of gene expression data using ANCut. *BMC Genomics*. 2017;18(1):623. [PubMed: 28814280]
25. Chalise P, Koestler DC, Bimali M, Yu Q, Fridley BL. Integrative clustering methods for high-dimensional molecular data. *Transl Cancer Res*. 2014;3(3):202–216. [PubMed: 25243110]
26. Li Y, Wu F-X, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform*. 2016;19(2):325–340.
27. Xing EP, Karp RM. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using norMalized cuts. *Bioinformatics*. 2001;17(suppl_1):S306–S315. [PubMed: 11473022]
28. Zhu R, Zhao Q, Zhao H, Ma S. Integrating multidimensional omics data for cancer outcome. *Biostatistics*. 2016;17(4):605–618. [PubMed: 26980320]
29. Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2000;22(8):888–905.
30. Nascimento MCV, De Carvalho AC. Spectral methods for graph clustering-a survey. *Eur J Oper Res*. 2011;211(2):221–231.
31. Zhang W, Feng H, Wu H, Zheng X. Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics*. 2017;33(17):2651–2657. [PubMed: 28472248]
32. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*. 2016;534:55–62. [PubMed: 27251275]

33. Yip PPC, Pao Y-H. Combinatorial optimization with use of guided evolutionary simulated annealing. *IEEE Trans Neural Netw.* 1995;6(2):290–295. [PubMed: 18263313]
34. Bertsimas D, Tsitsiklis J. Simulated annealing. *Stat Sci.* 1993;8(1):10–15.
35. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell.* 1979;2:224–227.
36. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Stat Methodol.* 2001;63(2):411–423.
37. Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. *Adv Neural Inf Process Syst.* 2002;2:849–856.
38. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet.* 2000;25(1):25–29. [PubMed: 10802651]
39. Akbani R, Akdemir KC, Aksoy BA, et al. Genomic classification of cutaneous melanoma. *Cell.* 2015;161(7):1681–1696. [PubMed: 26091043]
40. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511:543–550. [PubMed: 25079552]
41. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* 2010;16(2):176–195. [PubMed: 20013308]
42. Arias-Castro E, Pu X. A simple approach to sparse clustering. *Comput Stat Data Anal.* 2017;105:217–228.
43. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann Appl Stat.* 2011;5(1):232–253. [PubMed: 22081779]
44. Meng C, Helm D, Frejno M, Kuster B. moCluster: identifying joint patterns across multiple omics data sets. *J Proteome Res.* 2015;15(3):755–765. [PubMed: 26653205]
45. Huo Z, Tseng G. Integrative sparse K-Means with overlapping group Lasso in genomic applications for disease subtype discovery. *Ann Appl Stat.* 2017;11(2):1011–1039. [PubMed: 28959370]

FIGURE 1.

Simulation: estimated weights for one replicate (under scenario S1 with $a = 1$) under assisted weighed AWNCut and SKmeans. CNAs, copy number alterations; GEs, gene expressions [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2.

Data analysis: estimated weights under AWNCut for both gene expressions (GEs) and gene expressions (CNAs)

TABLE 1

Simulation scenarios S1 and S2

	<i>a</i>	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
S1	0.5	0.275(0.029)	0.289(0.012)	0.327(0.007)	0.326(0.008)	0.428(0.138)	0.450(0.088)	0.356(0.040)
	1.0	0.257(0.030)	0.271(0.019)	0.324(0.009)	0.320(0.012)	0.393(0.123)	0.439(0.108)	0.322(0.043)
	1.5	0.216(0.017)	0.263(0.009)	0.316(0.014)	0.317(0.013)	0.379(0.207)	0.481(0.143)	0.315(0.089)
S2	0.5	0.305(0.011)	0.334(0.011)	0.336(0.011)	0.328(0.007)	0.493(0.071)	0.461(0.105)	0.455(0.060)
	1.0	0.313(0.008)	0.333(0.012)	0.336(0.011)	0.313(0.010)	0.463(0.123)	0.482(0.182)	0.482(0.118)
	1.5	0.297(0.014)	0.322(0.012)	0.333(0.008)	0.308(0.007)	0.495(0.152)	0.485(0.131)	0.428(0.053)

In each cell, mean (sd) of M_{error} based on 500 replicates.

TABLE 2

Data analysis: size of each cluster

	AWNCut	NCutXZ	WNCut	NCut	Kmeans	SKmeans	Spectral
SKCM							
Cluster 1	31	45	50	50	45	106	31
Cluster 2	45	40	41	46	47	220	54
Cluster 3	45	50	48	48	29	8	40
Cluster 4	53	48	51	45	24	27	48
Cluster 5	51	49	42	50	30	2	57
Cluster 6	55	45	50	38	108	2	38
Cluster 7	46	48	42	46	21	1	48
Cluster 8	41	42	43	44	63	1	51
LUAD							
Cluster 1	137	153	162	169	33	255	143
Cluster 2	187	175	175	158	175	56	198
Cluster 3	179	175	166	176	295	192	162