HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2017 September 26.

Published in final edited form as:

Ann Appl Stat. 2017 June ; 11(2): 1011–1039. doi:10.1214/17-AOAS1033.

Integrative Sparse *K*-Means With Overlapping Group Lasso in Genomic Applications for Disease Subtype Discovery

Zhiguang Huo[1] and **George Tseng**[1]

Department of Biostatistics, University of Pittsburgh, Pittsburgh, ennsylvania 15261, USA

Department of Biostatistics, Human Genetics, and Computational Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA

Abstract

Cancer subtypes discovery is the first step to deliver personalized medicine to cancer patients. With the accumulation of massive multi-level omics datasets and established biological knowledge databases, omics data integration with incorporation of rich existing biological knowledge is essential for deciphering a biological mechanism behind the complex diseases. In this manuscript, we propose an integrative sparse *K*-means (is-*K* means) approach to discover disease subtypes with the guidance of prior biological knowledge via sparse overlapping group lasso. An algorithm using an alternating direction method of multiplier (ADMM) will be applied for fast optimization. Simulation and three real applications in breast cancer and leukemia will be used to compare is-*K* means with existing methods and demonstrate its superior clustering accuracy, feature selection, functional annotation of detected molecular features and computing efficiency.

Keywords and phrases

Cancer subtype; omics integrative analysis; overlapping group lasso; admm

1. Introduction

While cancer has been thought to be a single type of disease, increasing evidence from modern transcriptomic studies have suggested that each specific cancer may consist of multiple subtypes, with different disease mechanisms, survival rates and treatment responses. Cancer subtypes have been extensively studied, including in leukemia [Golub et al. (1999)], lymphoma [Rosenwald et al. (2002)], glioblastoma [Parsons et al. (2008); Verhaak et al. (2010)], breast cancer [Lehmann et al. (2011); Parker et al. (2009)], colorectal cancer [Sadanandam et al. (2013)] and ovarian cancer [Tothill et al. (2008)]. These subtypes usually have strong clinical relevance since they show different outcome, and might be responsive to different treatments [Abramson et al. (2015)]. However, single cohort/single omics (e.g., transcriptome) analysis suffers from sample size limitation and reproducibility issues [Simon et al. (2003); Simon (2005); Domany (2014)]. Over the years, large amount of omics data are accumulated in public databases and depositories, for

[1]Supported by the National Institutes of Health (NIH [RO1CA190766]).

example, The Cancer Genome Atlas (TCGA) http://cancergenome.nih.gov, Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/, Sequence Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra, just to name a few. These datasets provided unprecedented opportunities to reveal cancer mechanisms via combining multiple cohorts or multiple-level omics data types (a.k.a. horizontal omics meta-analysis and vertical omics integrative analysis; see below) [Tseng, Ghosh and Feingold (2012)]. Omics integrative analysis has been found successful in many applications: (e.g., breast cancer [Koboldt et al. (2012)], stomach cancer [Bass et al. (2014)]). On the other hand, a tremendous amount of biological information has been accumulated in public databases. Proper usage of these prior information (e.g., pathway information and miRNA targeting gene database) can greatly guide the modeling of omics integrative analysis.

In the literature, researchers have applied various types of clustering methods for high-throughput experimental data (e.g., microarray) to identify novel disease subtypes. Popular methods include hierarchical clustering [Eisen et al. (1998)], *K*-means [Dudoit and Fridlyand (2002)], mixture model-based approaches [Xie, Pan and Shen (2008); McLachlan, Bean and Peel (2002)] and nonparametric approaches [Qin (2006)], for analysis of single transcriptomic study. Resampling and ensemble methods have been used to improve stability of the clustering analysis [Kim et al. (2009); Swift et al. (2004)] or to pursue tight clusters by leaving scattered samples that are different from major clusters [Tseng (2007); Tseng and Wong (2005); Maitra and Ramler (2009)]. Witten and Tibshirani (2010) proposed a sparse *K*-means algorithm that can effectively select gene features and perform sample clustering simultaneously. To extend single-study techniques towards integration of multiple omics data sets, Tseng, Ghosh and Feingold (2012) categorized omics data integration into two major types: (A) horizontal omics meta-analysis and (B) vertical omics integrative analysis. For horizontal meta-analysis, multiple studies of the same omics data type (e.g., transcriptome) from different cohorts are combined to increase sample size and statistical power, a strategy often used in differential expression analysis [Ramasamy et al. (2008)], pathway analysis [Shen and Tseng (2010)] or subtype discovery [Huo et al. (2016)]. In contrast, vertical integrative analysis aims to integrate multi-level omics data from the same patient cohort (e.g., gene expression data, genome-wide profiling of somatic mutation, DNA copy number, DNA methylation or microRNA expression from the same set of biological samples [Richardson, Tseng and Sun (2016)]). In this paper, we focus on vertical omics integrative analysis for disease subtype discovery. Several methods for this purpose have been proposed in the literature. Lock and Dunson (2013) fitted a finite Dirichlet mixture model to perform Bayesian consensus clustering that allows common clustering across omics types as well as omics-type-specific clustering. The model, however, does not perform proper feature selection, and thus is not suitable for high-dimensional omics data. Shen, Olshen and Ladanyi (2009) proposed a latent variable factor model (namely iCluster) to cluster cancer samples by integrating multi-omics data. The method does not incorporate prior biological knowledge and requires extensive computing due to EM algorithm with large matrix operation. We will use the popular iCluster method as the baseline method to compare in this paper.

The central question we ask in this paper is: “Can we identify cancer subtypes by simultaneously integrating multi-level omics datasets and/or utilizing existing biological

knowledge to increase accuracy and interpretation?" Several statistical challenges will arise when we attempt to achieve this goal: (1) If multi-level omics data are available for a given patient cohort, what kind of method is effective to achieve robust and accurate disease subtype detection via integrating multi-omics data? (2) Since only a small subset of intrinsic omics features are relevant to the disease subtype characterization, how can we perform effective feature selection in the high-dimensional integrative analysis? (3) With the rich biological information (e.g., targeted genes of each miRNA or potential cis-acting regulatory mechanism between copy number variation, methylation and gene expression), how can we fully utilize the prior information to guide feature selection and clustering? In this paper, we propose an integrative sparse *K*-means (IS-*K* means) approach by extending the sparse *K*-means algorithm with overlapping group lasso technique to accommodate the three goals described above. The lasso penalty in the sparse *K*-means method allows effective feature selection for clustering. In the literature, (nonoverlapping) group lasso [Yuan and Lin (2006)] has been developed in a regression setting to encourage features of the same group to be selected or excluded together. The approach, however, has two major drawbacks: (1) it does not allow sparsity within groups (i.e., a group of features are either all selected or all excluded), and (2) the penalty function does not allow overlapping groups. For the first issue, Simon et al. (2013) proposed a sparse group lasso with both an l_1 lasso penalty and a group lasso penalty to allow sparsity within groups while the approach does not allow overlapping groups. For the latter issue, overlapping group information from biological knowledge is frequently encountered in many applications. In genomic application, for example, the targeted genes of two miRNAs are often overlapped or two pathways may contain overlapping genes. Jacob, Obozin-ski and Vert (2009) proposed a duplication technique to allow overlapping groups in regression setting while the approach does not allow sparsity within groups. In this paper, we attempt to simultaneously overcome both aforementioned difficulties in a clustering setting, which brings optimization challenges beyond the duplication technique by Jacob, Obozinski and Vert (2009) and the sparse group lasso optimization by Simon et al. (2013). In our proposed IS-*K* means method, we will develop a novel reformulation of l_1 lasso penalty and overlapping group lasso penalty so that a fast optimization technique using alternating direction method of multiplier (ADMM) [Boyd et al. (2011)] can be applied (see Section 3.4.1).

The rest of the paper is structured as following. Section 2 gives a motivating example. Section 3 establishes the method and optimization procedure. Sections 4.1–4.3 comprehensively compares the proposed method with the popular iCluster method using simulation and two breast cancer applications on multilevel omics data. Section 4.4 provides another type of IS-*K* means application of pathway-guided clustering on single transcriptomic study. Section 5 includes the final conclusion and discussion.

2. Motivating example

Figure 1(A) shows a clustering result using single study sparse *K*-means (detailed algorithm see Section 3.1) on the mRNA, methylation and copy number variation (CNV) datasets separately from 770 samples in TCGA. As expected, they generate very different disease subtyping without regulatory inference across mRNA, methylation and CNV. In this example, single study sparse *K*-means fails to consider that different omics features

belonging to the same genes are likely to contain cis-acting regulatory mechanisms related to the disease subtypes. Figure 1(B) combines the three datasets to perform *IS-K*means. The IS-*K* means generates a single disease subtyping and takes into account of the prior regulatory knowledge between mRNA, methylation and CNV. The prior knowledge can also be a pathway database (e.g., KEGG, BioCarta and Reactome) or knowledge of miRNA targets prediction databases (e.g., PicTar, TargetScan, DIANA-microT, miRanda, rna22 and PITA) [Witkos, Koscianska and Krzyzosiak (2011); Fan and Kurgan (2015)]. Incorporating such prior information of feature grouping increases statistical power and interpretation. Figure 1(C) shows a simple example of such group prior knowledge. Pathway $\mathscr{T}_1$ includes mRNA1, mRNA2, mRNA3 and mRNA6 while pathway $\mathscr{T}_2$ includes mRNA3, mRNA4, mRNA5 and mRNA7. Note that mRNA3 appears in both pathway $\mathscr{T}_1$ and $\mathscr{T}_2$, which requires our algorithm to allow overlapping groups. Our goal is to develop a sparse clustering algorithm integrating multi-level omics datasets and the aforementioned prior regulatory knowledge by overlapping group lasso. The algorithm is also suitable for single omics dataset with incorporating prior overlapping pathway information (see the leukemia examples in Section 4.4).

3. Method

3.1. K-Means and sparse K-means

Consider X_{jq} the gene expression intensity of gene j and sample q. The K-means method [MacQueen (1967)] targets to minimize the within-cluster sum of squares (WCSS):

$$\min_C \sum_{j=1}^{J} \mathrm{WCSS}_j(C) = \min_C \sum_{j=1}^{J} \sum_{K=1}^{K} \frac{1}{n_k} \sum_{p,q \in C_k} d_{pq,j}, \quad (3.1)$$

where K is the number of clusters, J is the number of genes (features), $C = (C_1, C_2, \ldots, C_k)$ denotes the clustering result containing partitions of all samples into K clusters, n_k is the number of samples in cluster k and $d_{pq,j} = (X_{jp} - X_{jq})^2$ denotes the squared Euclidean distance of gene j between sample p and q. One drawback of K-means is that it assumes all J features with equal weights in the distance calculation. In genomic applications, J is usually large but biologically only a small subset of genes may contribute to the sample clustering. Witten and Tibshirani (2010) tackled this problem by proposing a sparse K-means approach with lasso regularization on gene-specific weights. They found that direct application of lasso regularization to equation (3.1) will result in a meaningless null solution. Instead, they utilized the fact that minimizing *WCSS* is equivalent to maximizing between-cluster sum of squares (*BCSS*) since *WCSS* and *BCSS* add up to a constant value of total sum of squares [$TSS_j = BCSS_j(C) + WCSS_j(C)$]. The optimization in equation (3.1) is equivalent to

$$\max_C \sum_{j=1}^{J} \mathrm{BCSS}_j(C) = \max_C \sum_{j=1}^{J} \left[\frac{1}{n} \sum_{p,q} d_{pq,j} - \sum_{k=1}^{k} \frac{1}{n_k} \sum_{p,q \in C_k} d_{pq,j} \right]. \quad (3.2)$$

The lasso regularization on gene-specific weights in equation (3.2) gives the following sparse *K*-means objective function:

$$\max_{C,\mathbf{z}} \sum_{j=1}^{J} z_j \mathrm{BCSS}_j(C) \quad \text{subject to } \|\mathbf{z}\|_2 \le 1, \|\mathbf{z}\|_1 \le \mu, z_j \ge 0, \forall j, \quad (3.3)$$

where z_j denotes weight for gene *j*, $C = (C_1,\ldots, C_k)$ is the clustering result, *K* is the pre-estimated number of clusters and $\|\mathbf{z}\|_1$ and $\|\mathbf{z}\|_2$ arethe l_1 and l_2 norm of the weight vector $\mathbf{z} = (z_1,\ldots,z_J)$. The regularization shrinks most gene weights to zero and μ is a tuning parameter to control the number of nonzero weights (i.e., the number of intrinsic genes for subtype characterization). This objective function can be rewritten in its Lagrangian form:

$$\min_{C,\mathbf{z}} - \sum_{j=1}^{J} z_j \mathrm{BCSS}_j(C) + \gamma \|\mathbf{z}\|_1 \quad \text{subject to } \|\mathbf{z}\|_2 \le 1, z_j \ge 0, \forall J,$$

3.2. Integrative sparse K-means (IS-K means)

We extend the sparse *K*-means objective function to group structured sparse *K*-means. Here, we consider *J* to be the total number of features combing all levels of omics datasets. In order to make features of different omics data types on the same scale and comparable, we normalized $BCSS_j$ by TSS_j and denote

$$R_j(C) = \frac{\mathrm{BCSS}_j(C)}{\mathrm{TSS}_j}.$$

We put the overlapping group lasso penalty term $\Omega(\mathbf{z})$ in the objective function:

$$\min_{C,\mathbf{z}} - \sum_{j=1}^{J} z_j R_j(C) + \gamma\alpha \|\mathbf{z}\|_1 + \gamma(1-\alpha)\Omega(\mathbf{z}) \quad \text{subject to } \|\mathbf{z}\|_2 \le 1, z_j \ge 0, \quad (3.4)$$

where γ is the penalty tuning parameter controlling the numbers of nonzero features, $\alpha \in [0, 1]$ is a term controlling the balance between individual feature penalty and group feature penalty. If $\alpha = 1$, there is no group feature penalty term and the objective function is

equivalent to sparse K-means objective function after standardizing each feature. If $\alpha = 0$, there is no individual feature penalty and only group feature penalty exists. The overlapping group lasso penalty term is defined as

$$\Omega(\mathbf{z}) = \sum_{1 \le g \le \mathscr{G}_0} w_g \|\mathbf{m}_g \circ \mathbf{z}\|_2,$$

where $\mathscr{G}_0$ is the number of (possibly overlapping) feature groups from prior biological knowledge, $w_g \in \mathbb{R}$ is the group weight coefficient for group g, $\mathbf{m}_g = (\mathbf{m}_{g1}, \ldots, \mathbf{m}_{gJ})$ is the design vector of the gth feature group and $\circ$ represents Hadamard product. The design of w_g and $\mathbf{m}_g$ is discussed in Section 3.3. Note that features with no group information are also treated as a group by itself (a group only contains a feature); such a design is to avoid bias towards a feature with no group information by receiving no penalization. The feature groups can either come from existing biological databases (e.g., pathway or miRNA target database), or from basic biological cis-regulatory knowledge (CNV and methylation features in the neighborhood of a nearby gene region). The first term in equation (3.4) encourages large weights for features with strong clustering separability. The second term is an l_1 norm lasso penalty to encourage sparsity. Finally, $\Omega(\mathbf{z})$ serves as overlapping group lasso to encourage features in the prior knowledge groups to be selected simultaneously (or discarded together). The intuition of group lasso is that if we transform the Lagrange form of $\Omega(\mathbf{z})$ to its constraint form, it becomes an elliptic constraint and features of the same group are preferred to be selected together [Yuan and Lin (2006); Jacob, Obozinski and Vert (2009)]. The combination of l_1 norm lasso penalty and overlapping group lasso penalty $\Omega(\mathbf{z})$ serves to achieve a sparse feature selection and also encourages (but does not force) features of the same group to be selected together.

Remark. Since different types of omics datasets may have different value ranges and distributions, additional normalization may be needed in the preprocessing. For example, the commonly-used beta values from methy-seq (defined as “methylation counts”/“total counts”) represent the proportions of methylation and range between 0 and 1. A logit transformation to so-called M-values is closer to Gaussian distribution and is more suitable to integrate with other omics data. Similarly, log-transformation of expression intensities from microarray, log-transformation of RPKM/TPM (summarized expression values) from RNA-seq and log-ratio values of CNV values from SNP arrays have been shown to be roughly Gaussian distributed and are proper for multi-omics integration. Another possibility is by replacing Euclidean distance to an appropriate distance measurement (e.g., Gower's distance for binary categorical and ordinal data, and Bray– Curtis dissimilarity for count data). Under this scenario, equation (3.4) remains valid under such modification and we only need to incorporate partition around medoids (PAM) [Kaufman and Rousseeuw (1987)] instead of K-means in the optimization procedure in Section 3.4.1. However, heterogeneity of different distance measurement may require extra different sparsity penalties and this is beyond consideration in this paper.

3.3. Design of overlapping group lasso penalty

In this section, we discuss and justify the design of overlapping group lasso penalty for w_g and $\mathbf{m}_g$. We denote by $\mathcal{I}_g$ as the collection of features in group g ($1 \le g \le \mathcal{G}_0$) and define frequency of feature j appearing in different groups: $h(j) = \Sigma_{1 \le g \le \mathcal{G}_0} \mathbb{I}\{j \in \mathcal{I}_g\}$. We also define the intrinsic feature set $\mathcal{A}$ (i.e., features that contribute to the underlying true sample clustering) and the nonintrinsic feature set $\bar{\mathcal{A}}$. We first state an "unbiased feature selection" principle under a simplified situation:

Definition 3.1 ("Unbiased Feature Selection" principle). Suppose equal separation ability in all intrinsic features $\mathcal{A} = \{j : R_j = R > 0\}$ and no separation ability in nonintrinsic features $\bar{\mathcal{A}} = \{j : R_j = 0\}$ under the true clustering label. The proposed overlapping group lasso design (w_g and $\mathbf{m}_g$) is said to satisfy the "unbiased feature selection" principle if under equation (3.4), it generates equal weights $z_j = 1/\sqrt{|\mathcal{A}|}$ for $j \in \mathcal{A}$ and $z_j = 0$ for $j \in \bar{\mathcal{A}}$ given any prior knowledge of feature groups $\mathcal{I}_g$, $1 \le g \le \mathcal{G}_0$.

The theorem below states an overlapping group lasso penalty design that satisfies the "unbiased feature selection" principle when all features are intrinsic features (i.e., $\bar{\mathcal{A}} = (\phi)$.

Theorem 3.1. Consider $\Omega(\mathbf{z}) = \sum_{1 \le g \le \mathcal{G}_0} w_g \|\mathbf{m}_g \circ \mathbf{z}\|_2$ and $\mathbf{m}_g = (\mathbf{m}_{g1}, \ldots, \mathbf{m}_{gj}, \ldots, \mathbf{m}_{gJ})$ in equation (3.4). Suppose equal separation ability for all features $R_1 = \ldots = R_J = R$ ($\bar{\mathcal{A}} = \phi$) and further assume $R > \gamma$. The design of $\mathbf{m}_{gj} = \{j \in \mathcal{J}_g\} / \sqrt{h(j)}$, $w_g = \sqrt{\sum_{j \in \mathcal{J}_g} 1/h(j)}$ satisfies the "unbiased feature selection" principle such that optimum solution of $\mathbf{z}$ from equation (3.4) generates $z_j = 1/\sqrt{J}$, $\forall j$.

Theorem 3.1 gives a design of overlapping group lasso penalty such that given equal separation ability for all features, the feature selection is not biased by the prior group knowledge. When all the groups are nonoverlapping, $h(j) = 1$, $\forall j$, then

$$\Omega(\mathbf{z}) = \sum_{0 \le g \le \mathcal{G}_0} \left(\sqrt{|\mathcal{J}_g|} \sqrt{\sum_{j \in \mathcal{J}_g} z_j^2} \right),$$

where $|\mathcal{I}_g|$ is number of features in group $\mathcal{I}_g$, which is the nonoverlapping group lasso penalty [Yuan and Lin (2006)]. However, this weight design $(w_g = \sqrt{\sum_{j \in \mathcal{J}_g} 1/h(j)})$ is not applicable when the underlying intrinsic feature set is sparse (i.e., $\bar{\mathcal{A}} \neq \phi$). If there are many nonintrinsic features inside group g, the intrinsic features in group g is over penalized since w_g is inflated by the contribution of nonintrinsic features. Therefore, we propose the following overlapping group lasso penalty and show that the design satisfies the "unbiased feature selection" principle when the intrinsic feature set is sparse:

$$\mathbf{m}_{gj}=\{j \in \mathcal{J}_g\}/\sqrt{h(j)}, \quad w_g=\sqrt{\sum_{j\in(\mathcal{J}_g\cap\mathcal{I})} 1/h(j).} \tag{3.5}$$

Theorem 3.2. Suppose the intrinsic feature set $\mathcal{I} = \{j : R_j = R > 0\}$ and the nonintrinsic feature set $\bar{\mathcal{I}} = \{j : R_j = 0\}$. We further assume $R > \gamma$. The overlapping group lasso penalty in equation (3.5) satisfies the "unbiased feature selection" principle such that the optimum solution of $\mathbf{z}$ from equation (3.4) is $z_j = 1/\sqrt{|\mathcal{I}|}$ for $j \in \mathcal{I}$ and $z_j = 0$ for $j \in \bar{\mathcal{I}}$.

Note that we take into account both the nonintrinsic features and the intrinsic features in the penalty design in equation (3.5). Only intrinsic features contribute to the group weight coefficient w_g. The design vector $\mathbf{m}_g$ remains the same as nonoverlapping group lasso. In practice, the intrinsic feature set $\mathcal{I}$ is unknown. We follow the coefficient design of adaptive lasso [Zou (2006)] and adaptive group lasso [Huang, Horowitz and Wei (2010)], which have been discussed in the literature and they maintain a consistency property under certain mild conditions. Specifically, we set $\alpha = 1$ in equation (3.4) where only individual feature penalty is considered and use the solution $\hat{\mathbf{z}}$ to define estimated intrinsic feature set $\hat{\mathcal{I}} = \{j : \hat{\mathbf{z}}_j > 0\}$ and nonintrinsic feature set $\bar{\mathcal{I}} = \{j : \hat{\mathbf{z}}_j = 0\}$ for equation (3.5). In the example of Figure 1(C), suppose all 7 features are intrinsic genes. Pathway $\mathcal{T}_1$ contains mRNA1, mRNA2, mRNA3 and mRNA6, reflecting prior knowledge from pathway databases. Similarly, group for pathway $\mathcal{T}_2$ contains mRNA3, mRNA4, mRNA5 and mRNA7. As a result, $\mathbf{m}_1$ = (1, 1,1/2, 0, 0,1,0) and $\mathbf{m}_2$ = (0,0, 1/2,1,1, 0, 1) and

$$\Omega(z)=\sqrt{1+1+1/2+1}\sqrt{z_1^2+z_2^2+1/2\times z_3^2+z_6^2}+\sqrt{1/2+1+1+1}\sqrt{1/2\times z_3^2+z_4^2+z_5^2+z_7^2}.$$

Note that in our example mRNA3 is shared by pathway groups $\mathcal{T}_1$ and $\mathcal{T}_2$, representing overlapping group lasso penalty.

3.4. Optimization

In this section, we discuss major issues for optimization of equation (3.4). First, we introduce transformation of equation (3.4) such that l_1 norm penalty can be absorbed in l_2 norm group penalty. Second, we introduce the optimization procedure for the proposed objective function. Third, we discuss how to use ADMM to optimize the weight term, which is critical and a difficult problem since it involves both the l_1 norm penalty and overlapping group lasso penalty. Last, we discuss the stopping rule for the optimization.

3.4.1. Reformulation and iterative optimization—We use the fact that $\gamma\alpha\|\mathbf{z}\|_1$ can be rewritten as $\gamma\alpha\|\mathbf{z}\|_1=\gamma\alpha\sum_{j=1}^{J}\|\mathbf{z}_j\|_2$ and $\mathbf{z_j} = (0,\ldots, z_j, \ldots, 0)^\top$ with only the jth element nonzero. In other words, the l_1 norm penalty of a single feature can be deemed as group penalty with only one feature within a group. Therefore, we can rewrite objective function equation (3.4) as

$$\min_{C,z} -\sum_{j=1}^{J} z_j R_j(C) + \sum_{j=1}^{J} \|\gamma\alpha\boldsymbol{\phi}_j \circ \mathbf{z}\|_2 + \sum_{0\le g\le \mathscr{G}_0} \|\gamma(1-\alpha)\mathbf{m}_g \circ \mathbf{z}\|_2 \quad (3.6)$$

s.t. $\|\mathbf{z}\|_2 \le 1$, $z_j \ge 0$, where $\boldsymbol{\phi}_j = (\boldsymbol{\phi}_{j1}, \ldots, \boldsymbol{\phi}_{jJ})$, $\boldsymbol{\phi}_{ji} = 1$ if $j = i$ and $\boldsymbol{\phi}_{ji} = 0$ if $j \ne i$. We combine J and $\mathscr{G}_0$ groups and the combined groups are of size $\mathscr{G} = J + \mathscr{G}_0$. Define

$$\boldsymbol{\beta}_g = \begin{cases} \gamma\alpha\boldsymbol{\phi}_j & \text{if } 1 \le g \le J, \\ \gamma(1-\alpha)\mathbf{m}_g & \text{if } J+1 \le g \le \mathscr{G}. \end{cases}$$

Therefore, we can rewrite objective function equation (3.6) as

$$\begin{aligned} &\min -\mathbf{R}(C)^\top Z + \sum_{1\le g\le \mathscr{G}} \|\boldsymbol{\beta}_g \circ \mathbf{z}\|_2 \\ &\text{subject to } \|\mathbf{z}\|_2 \le 1, Z_j \ge 0, \end{aligned} \quad (3.7)$$

where $\mathbf{R}(C) = (R_1(C), \ldots, R_J(C))^\top$. The optimization procedure are outlined below:

1. Initialize weight **z** using the original sparse *K*-means method without the group lasso term.
2. Given weight **z**, use weighted *K*-means to update cluster labels C [**R** is the normalized WCSS so minimizing $-\mathbf{R}(C)^\top\mathbf{z}$ is essentially weighed *K*-means]. This is a nonconvex problem so multiple random starts are recommended to alleviate local minimum problem.
3. Given the cluster label C, **R** is fixed so optimizing the objective function is a convex problem with respect to solving weight **z**. We use ADMM in the next subsection to update weight **z**.
4. Iterate 2 and 3 until converge.

The detailed algorithm for Step 3 is outlined in Section 3.4.2 and the stopping rules of Step 3 and Step 4 are described in Section 3.4.3.

3.4.2. Update weight using ADMM—Alternating direction method of multiplier (ADMM) [Boyd et al. (2011)] is ideal for solving the optimization in equation (3.7). We introduce an auxiliary variable $\mathbf{x}_g$ and write down the augmented Lagrange:

$$\min -\mathbf{R}(C)^\top\mathbf{z} + \sum_{1\le g\le \mathscr{G}} \|\mathbf{x}_g\|_2 + \sum_{1\le g\le \mathscr{G}} \left\{ \mathbf{y}_g^\top(\mathbf{x}_g - \boldsymbol{\beta}_g \circ \mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}_g - \boldsymbol{\beta}_g \circ \mathbf{z}\|_2^2 \right\} \quad (3.8)$$

s.t. $\|\mathbf{z}\|_2 \le 1$, $z_j \ge 0$ and $\mathbf{x}_g = \boldsymbol{\beta}_g \circ \mathbf{z}$. This problem [equation (3.8)] is clearly equivalent to the original objective function [equation (3.7)], since for any feasible **z** the terms added to the objective is zero. ρ is the augmented Lagrange parameter which will be discussed in more detail in Section 3.4.4. Here, the augmented Lagrange is minimized jointly with respect to the two primal variables $\mathbf{x}_g$, **z** and the dual variable $\mathbf{y}_g$. In ADMM, $\mathbf{x}_g$, **z** and $\mathbf{y}_g$ are updated in an alternating or sequential fashion [Boyd et al. (2011)], and thus the optimization problem can be decomposed into three parts. Given ($\mathbf{X}_g^+$, **z** and $\mathbf{Y}_g^+$), the new iteration of (**x** +, $\mathbf{z}^+$ and $\mathbf{y}^+_g$) in equation (3.8) is updated as in the following:

$$\begin{cases} \mathbf{x}_g^+ = \arg\min_{\mathbf{x}_g} \|\mathbf{x}_g\|_2 + \mathbf{y}_g^\top \mathbf{x}_g + \frac{\rho}{2}\|\mathbf{x}_g - \boldsymbol{\beta}_g o \mathbf{z}\|_2^2, \\ \mathbf{z}^+ = \arg\min_{\mathbf{z}} -\sum z_j R_j - \sum_{1\le g\le \mathscr{G}} \mathbf{y}_g^\top (\boldsymbol{\beta}_g o \mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}_g^+ - \boldsymbol{\beta}_g o \mathbf{z}\|_2^2 \\ \qquad\qquad \text{subject to } \|\mathbf{z}\|_2 \le 1, z_j \ge 0, \\ \mathbf{y}_g^+ = \mathbf{y}_g + \rho(\mathbf{x}_g^+ - \boldsymbol{\beta}_g o \mathbf{z}^+), \end{cases}$$

where the updating equation of $\mathbf{x}_g^+$ and $\mathbf{z}^+$ are derived from equation (3.8) and the the updating equation of $\mathbf{y}_g^+$ is imbedded in ADMM procedure [Boyd et al. (2011)]. We can derive close form solution for $\mathbf{x}_g$ part and **z** part by the Karush– Kuhn–Tucker (KKT) condition. Details are given in the Appendix:

1. Define $\mathbf{a}_g = \boldsymbol{\beta}_g o \mathbf{z} - \frac{\mathbf{y}_g}{\rho}$ we have $\mathbf{x}_g^+ = (1 - \frac{1}{\rho\|\mathbf{a}_g\|_2}) + \mathbf{a}_g$ where $(\cdot)_+ = \max(0, \cdot)$.
2. Define $b_j = \sum_{1\le g\le \mathscr{G}} \rho \boldsymbol{\beta}_{gj}^2$ and $c_j = \sum_{1\le g\le \mathscr{G}} (\rho \mathbf{x}_{gj}^+ + \mathbf{y}_{gj}) o \boldsymbol{\beta}_{gj}$, where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g1}, \boldsymbol{\beta}_{g2}, \ldots, \boldsymbol{\beta}_{gJ})^\top$, $\mathbf{x}_g = (\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gJ})^\top$ and $\mathbf{y}_g = (\mathbf{y}_{g1}, \mathbf{y}_{g2}, \ldots, \mathbf{y}_{gJ})^\top$. The solution is given as following: we define $f_j(u) = (\frac{R_j + c_j}{b_j + 2u}) +$ If $\Sigma_j f_j(u)^2 < 1$, $z_j^+ = f_j(0) \forall j$ Otherwise, $z_j^+ = f_j(u) \forall j$ and u is selected s.t. $\|\mathbf{z}^+\|_2 = 1$.

3.4.3. Stopping rules—We have two algorithms which require stopping rules. For ADMM in the optimization of Step 3, the primal residual of group g in ADMM iteration t is: $\mathbf{r}t = \boldsymbol{x}^t - \boldsymbol{\beta}_g \circ \mathbf{z}^t$, and the l_2 norm of primal residual is $r^t = \sqrt{\sum_g \|\mathbf{r}_g^t\|_2^2}$. The l_2 norm of dual residual is: $\nu^t = \sqrt{\sum_g \|\boldsymbol{\beta}_g \circ (\mathbf{z}^t - \mathbf{z}^{t-1})\|_2^2}$. We set our ADMM stopping criteria such that simultaneously $r^t < 10^{-10}$ and $\nu^t < 10^{-10}$. For convergence of IS-K means, we iterate weighted K-means (Step 2) and updating weight by ADMM (Step 3) until converge. (i.e., $\frac{\sum_{j=1}^J |z_j^{(c)} - z_j^{(c-1)}|}{\sum_{j=1}^J |z_j^{(c-1)}|} < 10^{-4}$), where $z_j^{(c)}$ represents the z_j estimate in the c^{th} iteration of the IS-K means algorithm.

3.4.4. Augmented Lagrangian parameter ρ—Augmented Lagrangian parameter ρ controls the convergence of ADMM. In fact, large value of ρ will lead tosmall primal

residual by placing a large penalty on violations of primal feasibility. And conversely, small value of ρ tend to produce small dual residual, but it will result in a large primal residual by reducing the penalty on primal feasibility [Boydet al. (2011)]. An adaptive scheme of varying ρ to balance the primal and dualresidual has been proposed [He, Yang and Wang (2000); Wang and Liao (2001)]

which greatly accelerates ADMM convergence in practice:

$$\rho^{t+1} = \begin{cases} \tau^{\text{incr}}\rho^t & \text{if } \|r^t\|_2 > \eta\|\nu^t\|_2, \\ \rho^t/\tau^{\text{decr}} & \text{if } \|\nu^t\|_2 > \eta\|r^t\|_2, \\ \rho^t & \text{otherwise.} \end{cases}$$

We set $\eta = 10$ and $\tau^{\text{incr}} = \tau^{\text{decr}} = 2$. The intuition behind this scheme is to control both primal and dual residuals for converging to zero simultaneously.

3.5. Select tuning parameters

In the objective function of IS-K means, the number of clusters K is pre-specified. The issue of estimating K has been widely discussed in the literature and has been well recognized as a difficult and data-dependent problem. [Milligan and Cooper (1985); Kaufman and Rousseeuw (1990)]. Here, we suggest the number of clusters to be estimated in each study separately using conventional methods such as prediction strength [Tibshirani and Walther (2005)] or gap statistics [Tibshirani, Walther and Hastie (2001)] and jointly compared across studies (such that the numbers of clusters are roughly the same for all studies) for a final decision before applying integrative sparse K-means. Below we assume that a common K is pre-estimated for all omics datasets.

Another important parameter to be determined is α, which controls the balance between individual feature penalty and overlapping group penalty. According to equation (3.4), $\alpha = 1$ means we only emphasize on individual feature penalty and ignore overlapping group penalty. In this case, the IS-K means is equivalent to sparse K-means. $\alpha = 0$ means we only emphasize the overlapping group penalty and ignore the individual feature penalty. Simon et al. (2013) argued that there is no theoretically optimal selection for α because selection of α relates to multiple factors such as accuracy of prior group information and sparsity within groups. In general, a large α (e.g., $\alpha = 0.95$) is suitable when prior group information may not be accurate or features within selected groups may be sparse. On the other hand, if we expect mild sparsity within groups and high accuracy of prior group information, a small α (e.g., $\alpha = 0.05$) help select features by groups. In Section 4.1.2, we have performed simulation of different level of prior group information accuracy ($\theta = 1$ and $\theta = 0.2$) and found that $\alpha = 0.5$ generates robust and high performance results in the sensitivity analysis. As a result, we apply $\alpha = 0.5$ throughout the paper unless otherwise indicated.

The last tuning parameter is γ, which is the penalty coefficient. When γ is large, we place large penalty on the objective function and end up with less selected features. When γ is small, we put a small penalty and will include more features. We follow and extend the gap statistic procedure [Tibshirani, Walther and Hastie (2001)] to estimate γ:

1. For each feature in each omics type, randomly permute the gene expression (permute samples). This creates a permuted data set $X^{(1)}$. Repeat for B times to generate $X^{(1)}$, $X^{(2)}$, …, $X^{(B)}$.

2. For each potential tuning parameter γ, compute the gap statistics as below:

$$\text{Gap}(\gamma)=O(\gamma) - \frac{1}{B}\sum_{b=1}^{B} O_b(\gamma), \quad (3.9)$$

where $O(\gamma)= -\sum_{j=1}^{J} z_j^* R_j(C^*)$ is from observed data, where $\mathbf{z}^*$, C^* are the min-imizer of the objective function in equation (3.4) given γ. $O_b(\gamma)$ is similar to $O(\gamma)$ but generated from permuted data $X^{(b)}$.

3. For a range of selections of γ, select γ^* such that the gap statistics in equation (3.9) is minimized.

Figure 2 shows an example of a simulated dataset that will be discussed in Section 4.1. In this example, we used α = 0.5 for IS-K means and the minimum gap statistics corresponded to 1778 genes, which is very close to the underlying truth 1800. The gap statistics for α = 0.05, 0.95,1 are plotted in supplementary materials [Huo and Tseng (2017), Figure S1] and they all provided adequate γ estimation. In practice, calculating gap statistics from a chain of γ can be performed efficiently by adopting a warm start for adjacent γ's. For example, after calculating $O(\gamma_1)$, the resulting weights can be used as an initial value for the next nearby $\gamma_2 = \gamma_1 + \Delta$ to calculate $O(\gamma_2)$ in the optimization iteration for fast convergence.

4. Result

We evaluated integrative sparse K-means (IS-K means) on simulation datasets in Section 4.1, multiple-level omics applications using breast cancer TCGA (combining mRNA expression, DNA methylation and copy number variation) and METABRIC (combining mRNA expression and copy number variation) examples in Section 4.2 and 4.3, and a pathway-guided single transcriptomic application in leukemia in Section 4.4. In the simulation, the underlying sample clusters and intrinsic feature set were known and we demonstrated the better performance of IS-K means compared to iCluster and sparse K-means by cluster accuracy, feature selection and computing time. For the TCGA and METABRIC application, the underlying true clustering and intrinsic feature set were not known. We evaluated the performance by clustering similarity using adjusted Rand index (ARI) [Hubert and Arabie (1985)] with subtype definition by PAM50 [Parker et al. (2009)], cis-regulatory groups, survival difference between clusters and computing time. In the leukemia examples, the disease subtypes were defined by observable fusion gene aberration. We evaluated the performance by clustering accuracy (ARI) and pathway enrichment analysis on selected genes.

4.1. Simulation

4.1.1. Simulation setting—To assess the performance of integrative sparse *K*-means with different choices of α and compare to the original sparse *K* -means and iCluster, we simulated $K = 3$ subtypes characterized by several groups of subtype predictive genes in each of $S = 2$ omics datasets with $1 \le s \le S$ as the omics dataset index (e.g., $s = 1$ represents gene expression and $s = 2$ represents DNA methylation). The prior group information was imposed between groups of subtype predictive genes across omics datasets. These prior group information represent the possibility that a group of genes and DNA methylations might be co-regulated. To best preserve the data nature of genomic studies, we also simulated confounding variables, correlated gene structure and noninformative genes. Below is the generative process:

a. Subtype predictive genes (intrinsic feature set).

1. Denote by N_k is the number of subjects in subtype k ($1 \le k \le 3$). We simulate $N_1 \sim$ POI(40), $N_2 \sim$ POI(40), $N_3 \sim$ POI(30) and the number of subjects is $N = \Sigma_k N_k$. Simulate $S = 2$ omics datasets, which share the samples and subtypes. Specifically, we denote $s = 1$ to be the gene expression dataset and $s = 2$ to be the DNA methylation dataset.

2. Simulate $M = 30$ feature modules ($1 \le m \le M$) for each omics dataset. Denote n_{sm} to be the number of features in omics dataset s and module m. For each module in $s = 1$, sample $n_{1m} = 30$ genes. For each module in $s = 2$, sample $n_{2m} = 30$ methylations. Therefore, there will be of 1800 subtype predictive features among two omics datasets.

3. Denote by μ_{skm} is the template gene expression (on log scale) of omics dataset s ($1 \le s \le S$), subtype k ($1 \le k \le 3$) and module m ($1 \le m \le M$). Simulate the template gene expression $\mu_{skm} \sim N(9, 2^2)$ with constrain $\max_{p,q} |\mu_{spm} - \mu_{sqm}| \ge 1$, where p, q denote two subtypes. This part defines the subtype mean intensity for each module in all omics datasets. Note that since in equation (3.4) we used $R_j = \frac{\mathrm{BCSS}_j}{\mathrm{TSS}_j}$ for standardization, performance of the algorithm is robust to gene expression distribution (e.g., the Gaussian assumption here).

4. In order to tune the signal of the template gene expression, we introduce a relative effect size $f > 0$, such that $\mu'_{\mathrm{skm}} = (\mu_{\mathrm{skm}} - \min_k \mu_{\mathrm{skm}}) \times f + \min_k \mu_{\mathrm{skm}}$. If $f = 1$, we do not tune the signal. If $f < 1$, we decrease the signal and if $f > 1$, we amplify the signal.

5. Add biological variation $\sigma_1^2 = 1$ to the template gene expression and simulate $X'_{\mathrm{skmi}} \sim N(\mu'_{\mathrm{skm}}, \sigma_1^2)$ for each module m, subject i ($1 \le i \le N_k$) of subtype k and omics dataset s.

6. Simulate the covariance matrix Σ_{mks} for genes in module m, subtype k and omics dataset s, where $1 \le m \le M$, $1 \le k \le 3$ and $1 \le s \le S$. First

simulate $\sum'_{\mathrm{mks}} \sim \mathrm{W}^{-1}(\Phi, 100)$, where $\Phi = 0.5I_{nsm\times nsm} + 0.5J_{nsm\times nsm}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all elements equal 1. Then Σ_{mks} is calculated by standardizing $\sum'_{\mathrm{mks}}$ such that the diagonal elements are all 1 's.

7. Simulate gene expression levels of genes in cluster m as $(X_{1\mathrm{skmi}}, \ldots, X_{\mathrm{nsmskmi}})^{\top} \sim \mathrm{MVN}(X'_{\mathrm{skmi}}, \sum_{\mathrm{mks}})$, where $1 \le i \le N_{ks}$, $1 \le m \le M$, $1 \le k \le 3$ and $1 \le s \le S$.

b. Noninformative genes.

1. Simulate 5000 noninformative genes denoted by g ($1 \le g \le 5000$) in each omics dataset. First, we generate the mean template gene expression $\mu_{sg} \sim \mathrm{N}(9, 2^2)$. Then we add biological variance $\sigma_2^2=1$ to generate $X_{\mathrm{sgi}} \sim N(\mu_{sg}, \sigma_2^2), 1 \le i \le N_s$.

c. Confounder impacted genes.

1. Simulate $C = 2$ confounding variables. In practice, confounding variables can be gender, race, other demographic factors or disease stage etc. These will add heterogeneity to each study to complicate disease subtype discovery. For each confounding variable c ($1 \le c \le C$), we simulate $R = 10$ modules in each omics dataset. For each of these modules r_c ($1 \le r_c \le R$), sample number of genes $n_{r_c} = 30$. Therefore, totally 600 confounder impacted genes are generated in each omics dataset. This procedure is repeated in all S omics datasets.

2. For each omics dataset s ($1 \le s \le S$) and each confounding variable c, sample the number of confounder subclass $h_{sc} = k$. The N samples in omics dataset s will be randomly divided into h_{sc} subclasses.

3. Simulate confounding template gene expression $\mu_{slrc} \sim \mathrm{N}(9, 2^2)$ for confounder c, gene module r, subclass l ($1 \le l \le h_{sc}$) and omics dataset s. Similar to Step a5, we add biological variation $\sigma_1^2=1$ to the confounding template gene expression $X'_{\mathrm{scrli}} \sim N(\mu_{\mathrm{slrc}}, \sigma_1^2)$. Similar to Steps a6 and a7, we simulate gene correlation structure within modules of confounder impacted genes.

d. Gene grouping information.

1. We assume omics dataset $s = 1$ and $s = 2$ have prior group information on subtype predictive gene modules. There are $M = 30$ modules in each omics dataset.

2. Suppose subtype predictive genes in the mth module of the first omics dataset are grouped with methylation features in the second omics dataset (totally $n_{1m} + n_{2m} = 30 + 30 = 60$ features are in the same group). With probability $1 - \theta$ ($0 \le \theta \le 1$), each feature out of the 60

features will be randomly replaced by a confounder impacted gene or noninformative gene. Note that the same replaced feature can appear in multiple subtype predictive gene groups. We set $\theta = 1$ and 0.2 to reflect 100%, 20% accuracy of prior group information.

4.1.2. Simulation result—For IS-*K* means, the tuning parameter γ was selected by gap statistics introduced in Section 3.5. Table 1 shows the result of gap statistics to select the best γ in the simulation of $\alpha = 0.5$, $\theta = 1$. The smallest gap statistics was selected at $\gamma = 0.21$ that correspond to selecting 1778 features, which was close to the underlying truth. Similarly, gap statistics result for $\alpha = 1, 0.95, 0.05$ are in the supplementary materials [Huo and Tseng (2017), Figure S1]. For simulation, we generated two scenarios with relative effect size $f = 0.6$ and $f = 0.8$. The complete simulation result of $f = 0.6$ is shown in Table 1 and the result for $f = 0.8$ is in the supplementary materials [Huo and Tseng (2017), Table S1]. For iCluster and sparse *K*-means, we allowed them to choose their own optimum tuning parameters. Note that sparse *K*-means was adopted to each individual omics datatype. We used ARI [Hubert and Arabie (1985)] and Jaccard index [Jaccard (1901)] to evaluate the clustering and feature selection performance. ARI calculated similarity of the clustering result with the underlying true clustering in simulation (range from −1 to 1 and 1 represents exact same partition compared to the underlying truth). Jaccard index compared the similarity and diversity of two feature sets, defined as the size of the intersection of two feature sets divided by the size of the union of two feature sets (range from 0 to 1 and 1 represent identical feature sets compared to the underlying truth). Clearly, IS-*K* means outperformed iCluster and individual study sparse *K*-means in terms of ARI and Jaccard index. IS-*K* means and sparse *K*-means outperformed iCluster in terms of computing time. Within IS-*K* means, we compared feature selection in terms of area under the curve (AUC) of ROC curve, which avoids the issue of tuning parameter selection. When $\theta = 1$ (representing the grouping information is accurate), smaller α (representing larger emphasize on grouping information) yielded better feature selection performance in terms of AUC as expected. However, when $\theta = 0.2$ (representing many errors in the grouping information), smaller α yielded worse performance in terms of AUC. Note that $\alpha = 0.5$ gives robustness and performs well in the two extremes of $\theta = 1$ and $\theta = 0.2$. In all applications below, we will apply $\alpha = 0.5$ unless otherwise noted.

4.1.3. Data perturbation—We also evaluated the stability of the algorithm against data perturbation. Instead of Gaussian distribution in the data generative process, we utilized heavy tailed t-distribution to generate the expression. In the simulation setting Step a3, the template gene expression is simulated from a t-distribution with degree of freedom 3, location parameter 9 and scale parameter 2. In Step a4, we set relative effect size $f = 0.6$ and $f = 0.8$, respectively. In Step a5, X'_{skmi} is simulated from a t-distribution with degree of freedom 3, location parameter μ'_{skmi} s and scale parameter σ_1^2. The result for data perturbation is in supplementary materials [Huo and Tseng (2017), Tables S5 and S6]. The resulting message remains almost the same as the conclusion in Section 4.1.2. Therefore, our proposed algorithm is robust against non-Gaussian or heavy tail distributions.

4.2. Integrating TCGA breast cancer mRNA, CNV and methylation

We downloaded TCGA breast cancer (BRCA) multi-level omics datasets from TCGA NIH official website. TCGA BRCA gene expression (IlluminaHiSeq RNAseqV2) was downloaded on 04/03/2015 with 20,531 genes and 1095 subjects. TCGA BRCA DNA methylation (Methylation450) was downloaded on 09/12/2015 with 485,577 probes and 894 subjects. TCGA BRCA copy number variation (BI gis-tic2) was downloaded on 09/12/2015 with 24,776 genes and 1079 subjects. There were 770 subjects with all these three omics data types. Features (probes/genes) with any missing value were removed. For gene expression, we transformed the FPKM value by $\log_2(\cdot + 1)$, where 1 is a pseudo-count to avoid undefined $\log_2(0)$, such that the transformed value was on continuous scale. For methylation, the Methylation450 platform provided beta value with range $0 < \beta < 1$, where 0 represents un-methylated and 1 represents methylated. We transformed the beta value to M value, which is defined by a logit transformation $(M = \log_2[\frac{\beta}{1-\beta}])$. Therefore, methylation characterized by M value is on a continuous scale, similar to mRNA and CNV. If multiple methylation probes matched to the same gene symbol, we selected one methylation probe as a representative, which had the largest average correlation with other methylation probes of the same gene symbol. We ended up with 20,147 methylation probes with unique gene symbols.

We filtered out 50% low expression genes (unexpressed genes) and then 50% low variance genes (noninformative genes). 50% low expression genes are genes with the lowest 50% mean of gene expression across samples and 10,250 genes remained after this filtering step. 50% low variance genes are genes with the lowest 50% variance of gene expression across samples and 5125 genes remained after this filtering step. We obtained 4815 CNV features and 5035 methylation features by matching to the 5125 gene symbols. The features from three different omics datasets that shared the same cis-regulatory annotation (same gene symbol) were grouped together to form 5125 feature groups. In this case, each group had one mRNA gene expression, one CNV gene and/or one methylation probe. Each group contained candidate multi-omics regulatory information because CNV and methylation could potentially regulate mRNA expression. We applied IS-K means with $\alpha = 0.5$, sparse K-means by directly merging three omics datasets together as well as iCluster. Number of clusters K was set to be 5 since it was well established that breast cancer has 5 subtypes by PAM50 definition [Parker et al. (2009)]. For a fair comparison, we selected the tuning parameter for each method such that number of selected features are close to 2000.

For evaluation purposes, we investigated three categories of groups among selected features: G1, G2 and G3. G3 represents feature groups (gene symbol) where all three types (mRNA, CNV and methylation) of features are selected. Similarly, G2 represents feature groups (gene symbol) where only two types of features are selected; G1 represents feature groups (gene symbol) where only one type of feature is selected. We also compared the clustering result with PAM50 subtype definition in terms of ARI. The result is shown in Table 2. Clearly, IS-K means obtained more G2 and G3 features than sparse K-means and iCluster. This is biologically more interpretable but not surprising since IS-K means incorporated the multi-omics regulatory information and we expected feature of the same group were

encouraged to come out together. Besides, IS-K means has higher ARI compared to sparse K-means and iCluster, indicating the clustering result of IS-K means is closer to PAM50 definition than sparse K-means and iCluster. The 5-by-5 confusion table of IS-K means clustering result and PAM50 subtypes is shown in supplementary materials [Huo and Tseng (2017), Table S3]. One should note the the ARI for all these three methods are not very high. This could be because PAM50 was defined by gene expression only and in our scenario we integrated multi-omics information. The heatmaps of IS-K means result is shown in Figure 1(B). In terms of computing time, IS-K means is nearly 20 times faster than iCluster.

4.3. Integrating METABRIC breast cancer mRNA and CNV

We tested the performance of IS-K means in another large breast cancer multi-omics (sample size n = 1981) dataset METABRIC [Curtis et al. (2012)] with mRNA expression (llu-mina HumanHT12v3) and CNV (Affymetrix SNP 6.0 chip) and survival information. The datasets are available at https://www.synapse.org/#Synapse:syn1688369/wiki/27311. There were originally 49,576 probes in gene expression. If multiple probes matched to the same gene symbol, we selected the probe with the largest IQR (interquartile range) to represent the gene. After mapping the probes to gene symbols, we obtained 19,489 mRNA expression features and 18,538 CNV features, which shared 1981 samples. After filtering out 30% low expression mRNA based on mean gene expression across samples and then 30% low variance mRNA based on variance of gene expression across samples, we ended up with 9504 mRNA features. We obtained 8696 CNV feature symbols by matching with mRNA feature symbols. Therefore, we had totally 18,200 features and 9504 feature groups (share the same gene symbol) among 1981 samples.

We applied IS-K means with α = 0.5, sparse K-means by directly merging three omics dataset together as well as iCluster. The number of clusters K was set to be 5 (same reason in TCGA). For a fair comparison, we selected the tuning parameter for each method such that number of selected features are close to 2000. For evaluation purposes, we similarly defined two categories of groups among selected features. G2 represents feature groups (gene symbol) where both types of features are selected and G1 represents feature groups (gene symbol) where only one type of feature is selected. We also compared the clustering result with PAM50 subtype definition in terms of ARI. The result is shown in Table 3. Similar to the TCGA example in Section 4.2, IS-K means obtained more G2 features than sparse K-means and iCluster. The log-rank test of survival difference for the clustering result defined by IS-K means is more significant than sparse K-means and iClus-ter. Furthermore, IS-K means has higher ARI compared to sparse K-means and iCluster, indicating the clustering result of IS-K means is closer to PAM50 definition than sparse K-means and iCluster. The 5-by-5 confusion table of IS-K means clustering result and PAM50 subtypes are in the supplementary materials [Huo and Tseng (2017), Table S4]. In terms of computing time, IS-K means and sparse K-means are much faster than iCluster.

4.4. Three leukemia transcriptomic datasets using pathway database as prior knowledge

In the simulations and applications so far (Sections 4.1–4.3), we have focused on using the cis-regulatory mechanism as grouping information for integrating multi-level omics data for sample clustering. In this subsection, we present a different but commonly encountered

application of pathway-guided clustering in single transcriptomic study. Specifically, we use pathway information from databases to provide prior overlapping group information (i.e., a pathway is a group containing tens to hundreds of genes and two pathways may contain overlapping genes). A transcriptomic study is used for sample clustering with the overlapping group information. We apply IS-*K* means to three leukemia transcrip-tomic datasets [Verhaak et al. (2009); Balgobind et al. (2010) and Kohlmann et al. (2008)] separately and using three pathway databases (KEGG, BioCarta and Reac-tome) independently, generating nine IS-*K* means clustering results (see Table 4). The supplementary materials [Huo and Tseng (2017), Table S2] show a summary description of the three leukemia transcriptomic studies.

We only considered samples from acute myeloid leukemia (AML) with three fusion gene subtypes: inv(16) (inversions in chromosome 16), t(15; 17) (translo-cations between chromosome 15 and 17), t(8;21) (translocations between chromosomes 8 and 21). These three gene-translocation AML subtypes have been well studied with different survival, treatment response and prognosis outcomes. Since the three subtypes are observable under the microscope, we treated these class labels as the underlying truth to evaluate the clustering performance. The expression data for Verhaak, Balgobind ranged from around [3.169, 15.132] while Kohlmann ranged in [0, 1]. All the datasets were downloaded directly from the NCBI GEO website. Originally, there were 54,613 probe sets in each study. For each study, we removed genes with any missing value in it. If multiple microarray probes matched to the same gene symbol, we selected the probe with the largest interquartile range (IQR) to represent the gene. We ended up with 20,154 unique genes in Verhaak and 20,155 unique genes in Balgobind and Kohlmann. We further filtered out 30% low expression genes in each study, which were defined as 30% of genes with the lowest mean expression. We ended up with 14,108 unique genes in each study.

We obtained the three pathway databases (BioCarta, KEGG and Reactome) from MSigDB (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2) as the prior group information to guide feature selection in IS-*K* means. The original pathway sizes were 217, 186 and 674 for BioCarta, KEGG and Reactome. We only kept pathways with size (number of genes inside pathway) greater or equal to 15 and less or equal to 200 after intersecting with 14,108 unique genes. After gene size restriction, we ended up with 114, 160 and 428 pathways for BioCarta, KEGG and Reactome. Note that these pathway groups have large overlaps (i.e., many genes appear in multiple pathways).

For each of the three studies, we applied IS-*K* means (with BioCarta, KEGG and Reactome as prior group information, respectively), sparse *K*-means and iCluster. Note that in this example, IS-*K* means dealt with single omics dataset with prior knowledge. For a fair comparison, we tuned the parameters so that the number of selected features are close to 1000. The result is shown in Table 4. For Verhaak and Kohlmann, IS-*K* means and sparse *K*-means almost recovered the underlying true clustering labels (ARI = 0.901–0.932), while iCluster had relatively smaller ARI (ARI = 0.733). We investigated the heatmap of the clustering result of Verhaak using iCluster (supplementary materials [Huo and Tseng (2017), Figure S2]) to understand reasons of its worse performance (lower ARI) and found that its solution converged to a stable clustering configuration with clear clustering separation.

Thus, the worse clustering performance in iCluster likely comes from a local optimum solution. For Balgobind, the clustering results from IS-*K* means and sparse *K*-means had smaller ARI (ARI = 0.792) but iCluster performed even worse (ARI = 0.214).

To further evaluate functional annotation of the selected intrinsic genes via each method, we explored pathway enrichment analysis (Figure 3) using BioCarta database via Fisher exact test. Five methods [iCluster, IS-*K* means (BioCarta), IS-*K* means (KEGG), IS-*K* means (Reactome), sparse *K*-means] were compared. The jittered plot of $-\log_{10} p$-values is shown in Figure 3. IS-*K* means (BioCarta) show the most significant pathways consistently across three studies; this is somewhat expected since we used the BioCarta pathway as prior knowledge to guide our feature selection. IS-*K* means (KEGG) and IS-*K* means (Reactome) also showed more significant pathways than sparse *K*-means and iCluster, indicating incorporating prior knowledge indeed improved feature selection (in the sense that the selected feature are more biological meaningful). Note that IS-*K* means (KEGG) and IS-*K* means (Reactome) did not have an overfitting issue since the test pathway database (BioCarta) was different from the prior knowledge we utilized. Similarly, the results using KEGG and Reactome as a testing pathway are in supplementary materials [Huo and Tseng (2017), Figure S3].

5. Conclusion and discussion

Cancer subtype discovery is a critical step for personalized treatment of the disease. In the era of massive omics datasets and biological knowledge, how to effectively integrate omics datasets and/or incorporate existing biological evidence brings new statistical and computational challenges. In this paper, we proposed an integrative sparse *K*-means (IS-*K* means) approach for this purpose. The existing biological information is incorporated in the model and the resulting sparse features can be further used to characterize the cancer subtype properties in clinical application.

Our proposed IS-*K* means has the following advantages. First, integrative analysis increases clustering accuracy, statistical power and explainable regulatory flow between different omics types of data. The existing biological information is taken into account by using the overlapping group lasso. Fully utilizing the inter-omics regulatory information and external biological information will increase the accuracy and interpretation of the cancer subtype findings. Second, we reformulated the complex objective function into a simplified form where weighted *K*-means and ADMM can be iteratively applied to optimize the convex sub-problems with closed-form solutions. Due to the nature of classification EM algorithm in *K*-means and closed-form iteration updates of ADMM, implementation of the IS-*K* means framework is computationally efficient. IS-*K* means only takes 10-15 minutes for 15,000 omics features and more than 700 subjects on a standard desktop with single computing thread while iCluster takes almost 4 hours. Third, the resulting sparse features from IS-*K* means have better interpretation than features selected from iCluster.

IS-*K* means potentially has the following limitations. The existing biological information is prone to errors and can be updated frequently. Incorporating false biological information may dilute information contained in the data and even lead to biased finding. Therefore, we

suggest not to over-weigh the overlapping group lasso term and choose $\alpha = 0.5$ to adjust for the balance between information from existing biological knowledge and information from the omics datasets. The users can, however, tune this parameter depending on the strength of their prior belief of the biological knowledge. Another limitation is that IS-*K* means can only deal with one cohort with multiple types of omics data. How to effectively combine multiple cohorts with multi-level omics data will be a future work. R package "ISKmeans" incorporates C++ for fast computing and it is publicly available on GitHub https://github.com/Caleb-Huo/IS-Kmeans as well as the authors' websites. All the data and code presented in this paper are also available on the authors' websites.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors appreciate the many insightful critiques from the reviewers and Associate Editor.

Appendix

A.1. Proof for Theorem 3.1 and Theorem 3.2

Proof of Theorem 3.1. Given equal separation ability for each feature $R_1 = \ldots = R_j = \ldots = R_J = R$ and the proposed design of overlapping group lasso penalty, equation (3.4) becomes

$$\min_{C,\mathbf{z}} -\sum_{j=1}^{J}\mathbf{z}_j R+\gamma\alpha\|\mathbf{z}\|_1+\gamma(1-\alpha)\sum_{1\le g\le \mathscr{G}_0}\left(\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)}\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)\times z_j^2}\right) \quad \text{subject to } \|\mathbf{z}\|_2\le 1, z_j\ge 0, \forall j.$$

First, we can take away the constraint $z_j \geq 0$, $\forall j$. It is easy to see that if any $z_j < 0$, we can always use $-z_j$ to replace the solution and the objective function will decrease. We can write down the Lagrange function of equation (3.4) after dropping the constraint $z_j \geq 0$, $\forall j$:

$$L(\mathbf{z},\lambda)=-\sum_{j=1}^{J}z_j R+\gamma\alpha\|\mathbf{z}\|_1+\gamma(1-\alpha)\sum_{1\le g\le \mathscr{G}_0}\left(\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)}\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)\times z_j^2}\right)+\lambda(\|\mathbf{z}\|_2^2-1).$$

Partial derivative of the Lagrange is

$$\frac{\partial \boldsymbol{L}(\mathbf{z})}{\partial_{Z_j}}=-R+\gamma\alpha\frac{\partial|z_j|}{\partial z_j}\quad+\gamma(1-\alpha)\sum_{1\le g\le \mathscr{G}_0}\left(\sqrt{\sum_{j'\in\mathscr{J}_g}1/h(j')}\frac{\{j\in\mathscr{J}_g\}\times 1/h(j)\times Z_j}{\sqrt{\sum_{j'\in\mathscr{J}_g}1/h(j')\times z_{j'}^2}}\right)+2\lambda z_j.$$

It is easy to verify that $z_1=z_2=\cdots=z_j=1/\sqrt{J}, \lambda=\frac{\sqrt{J}(R-\gamma)}{2}$ will make $\frac{\partial L(\mathbf{z})}{\partial z_j}=0, \forall_j$. Since the object function is a convex function, according to sufficiency of the KKT condition, the proposed penalty design will lead to the solution of the "unbiased feature selection" principle.

Proof of Theorem 3.2. For the intrinsic gene set $\mathcal{G}$, we have $R_j = R > 0$ for $j \in \mathcal{G}$. For the nonintrinsic gene set $\bar{\mathcal{G}}$, we have $R_j = 0$ for $j \in \bar{\mathcal{G}}$. Given the proposed design of overlapping group lasso penalty, equation (3.4) becomes

$$\min_{C,\mathbf{z}} -\sum_{j=1}^{J}\mathbf{z}jR(j\in\mathscr{I})+\gamma\alpha\|\mathbf{z}\|_1+\gamma(1-\alpha)\sum_{1\le g\le\mathscr{G}_0}\left(\sqrt{\sum_{j\in(\mathscr{J}_g\cap\mathscr{I})}1/h(j)}\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)\times z_j^2}\right) \quad \text{subject to } \|\mathbf{z}\|_2\le 1, z_j\ge 0, \forall j.$$

First, we can similarly take away the constraint $z_j \geq 0$, $\forall j$. We can write down the Lagrange function of equation (3.4) after dropping the constraint $z_j \geq 0$, $\forall_j$:

$$\boldsymbol{L}(\mathbf{z},\lambda)=-\sum_{j=1}^{J}\mathbf{z}j\boldsymbol{R}(j\in\mathscr{I})+\gamma\alpha\|\mathbf{z}\|_1+\gamma(1-\alpha)\sum\nolimits_{1\le g\le\mathscr{G}_0}\left(\sqrt{\sum_{j\in(\mathscr{J}_g\cap\mathscr{I})}1/h(j)}\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)\times z_j^2}\right)+\lambda(\|\mathbf{z}\|_2^2-1).$$

The partial derivative of the Lagrange is

$$\frac{\partial\boldsymbol{L}(\mathbf{z})}{\partial_{z_j}}=-R(j\in\mathscr{I})+\gamma\alpha\frac{\partial|z_j|}{\partial z_j}+\gamma(1-\alpha)\sum_{1\le g\le\mathscr{G}_0}\left(\sqrt{\sum_{j'\in(\mathscr{J}_g\cap\mathscr{I})}1/h(j')}\frac{\{j\in\mathscr{J}_g\}\times 1/h(j)\times Z_j}{\sqrt{\sum_{j\in\mathscr{J}_g}1/h(j)\times z_j^2}}\right)+2\lambda z_j.$$

It is easy to verify that if for $j \in \mathcal{G}$, $z_j = 1/\sqrt{J}$, $j \in \bar{\mathcal{G}}$, $z_j = 0$ and $\lambda=\frac{\sqrt{J}(R-\gamma)}{2}$ is a zero solution to the partial derivative of the Lagrange function. Note here we set the sub-gradient $\frac{\partial|z_j|}{\partial z_j}=0$ at $z_j = 0$. Since the object function is a convex function, according to sufficiency of KKT condition, the proposed penalty design leads to the "unbiased feature selection" principle.

A.2. Optimization by KKT condition

There are two optimization problems:

$$\begin{cases} \mathbf{x}_g^+ = \arg\min_{\mathbf{x}_g} \|\mathbf{x}_g\|_2 + \mathbf{y}_g^\top \mathbf{x}_g + \frac{\rho}{2}\|\mathbf{x}_g - \beta_g \circ \mathbf{z}\|_2^2, \\ \mathbf{z}^+ = \arg\min_{\mathbf{z}} -\sum \mathbf{z}_j \boldsymbol{R}_j - \sum_{1\le g\le \mathscr{G}} \mathbf{y}_g^\top (\beta_g \circ \mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}_g^+ - \beta_g \circ z\|_2^2 \\ \text{subject to } \|\mathbf{z}\|_2 \le 1,\ Z_j \ge 0. \end{cases}$$

It is a convex optimization problem for $\mathbf{x}_g^+$ with no constraint. The stationarity condition states that the sub-gradient of the objective function will be 0 at the optimum solution. Therefore, we have

$$S(\mathbf{x}_g^+) + \mathbf{y}_g + \rho(\mathbf{x}_g^+ - \beta_g \circ \mathbf{z}) = 0,$$

where $S(\mathbf{v})$ is the sub-gradient of $\|\mathbf{v}\|_2$ and

$$S(\mathbf{v}) \in \begin{cases} \frac{\mathbf{v}}{\|\mathbf{v}\|_2} & \text{if } \|\mathbf{v}\|_2 \ge 1, \\ 0 & \text{otherwise.} \end{cases}$$

If we define $\mathbf{a}_g = \beta_g \circ \mathbf{z} - \frac{\mathbf{y}_s}{\rho}$, it can be derived that $\mathbf{x}_g^+ = (1 - \frac{1}{\rho\|\mathbf{a}_g\|_2})_+ \mathbf{a}_g$ where $(\cdot)_+ = \max(0, \cdot)$.

The optimization problem for $\mathbf{z}^+$ is a convex optimization problem with two constraints. We first write down the Lagrange function and convert the constrained optimization problem into an un-constrained optimization problem:

$$\arg\min_{\mathbf{z}} -\sum_j \mathbf{z}_j \boldsymbol{R}_j - \sum_{1\le g\le \mathscr{G}} \mathbf{y}_g^\top (\beta_g \circ \mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}_g^+ - \beta_g \circ \mathbf{z}\|_2^2 + u(\|\mathbf{z}\|_2 - 1) - \sum_j \nu_j z_j$$

such that $u \in \mathbb{R}$, $u \ge 0$, $\nu_j \in \mathbb{R}$ and $\nu_j \ge 0\ \forall_j$. Taking gradient of the Lagrange function with respect to $\mathbf{z}$ and use the constraints, we can derive the solution to this problem. Define $b_j = \sum_{1\le g\le \mathscr{G}} \rho \boldsymbol{\beta}_{gj}^2$ and $c_j = \sum_{1\le g\le \mathscr{G}} (\rho \mathbf{x}_{gi}^+ + \mathbf{y}_{gi} \circ \mathbf{m}_{gi})$, where $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g1}, \boldsymbol{\beta}_{g2}, \ldots, \boldsymbol{\beta}_{gJ})^\top$, $\mathbf{x}_g = (\mathbf{x}_{g1}, \mathbf{x}_{g2}, \ldots, \mathbf{x}_{gJ})^\top$, $\mathbf{y}_g = (\mathbf{y}_{g1}, \mathbf{y}_{g2}, \ldots, \mathbf{y}_{gJ})^\top$, and $\mathbf{m}_g = (\mathbf{m}_{g1}, \mathbf{m}_{g2}, \ldots, \mathbf{m}_{gJ})^\top$. The solution is given as following: we define $f_i(u) = (\frac{R_j + c_j}{b_j + 2u})_+ \cdot$ if $\sum_j f_i(u)^2 < 1$. If $\Sigma_j f_j(u)^2 < 1$, $z_j^+ = f_i(0)$. Otherwise, $z_j^+ = f_i(u)$ and u is selected s.t. $\|\mathbf{z}^+\|_2 = 1$.

References

Abramson VG, Lehmann BD, Ballinger TJ, Pietenpol JA. Subtyping of triple-negative breast cancer: Implications for therapy. Cancer. 2015; 121:8–16. [PubMed: 25043972]

Balgobind BV, Van den Heuvel-Eibrink MM, De Menezes RX, Reinhardt D, Hollink IHIM, Arentsen-Peters STJCM, van Wering ER, Kaspers GJL, Cloos J, de Bont ESJM, Cayuela JM, Baruchel A, Meyer C, Marschalek R, Trka J, Stary J, Beverloo HB, Pieters R, Zwaan CM, den Boer ML. Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. Haematologica. 2010; 96:221–230. [PubMed: 20971820]

Bass A, Thorsson V, Shmulevich I, Reynolds S, Miller M, Bernard B, Hinoue T, Laird P, Curtis C, Shen H, et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014; 513:202–209. [PubMed: 25079317]

Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn. 2011; 3:1–122.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486:346–352. [PubMed: 22522925]

Domany E. Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. Cancer Res. 2014; 74:4612–4621. [PubMed: 25183786]

Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. 2002; 3:1–21.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998; 95:14863–14868. [PubMed: 9843981]

Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. Brief Bioinform. 2015; 16:780–794. [PubMed: 25471818]

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–537. [PubMed: 10521349]

He BS, Yang H, Wang SL. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. J Optim Theory Appl. 2000; 106:337–356. MR1788928.

Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. Ann Statist. 2010; 38:2282–2313. MR2676890.

Hubert L, Arabie P. Comparing partitions. J Classification. 1985; 2:193–218.

Huo Z, Tseng G. Supplement to “Integrative sparse *K*-means with overlapping group lasso in genomic applications for disease subtype discovery”. 2017; doi: 10.1214/17-AOAS1033SUPP

Huo Z, Ding Y, Liu S, Oesterreich S, Tseng G. Meta-analytic framework for sparse *K*-means to identify disease subtypes in multiple transcriptomic studies. J Amer Statist Assoc. 2016; 111:27–42. MR3494636.

Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaud Sci Nat. 1901; 37:547–579.

Jacob, L., Obozinski, G., Vert, JP. ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. ACM; New York: 2009. Group lasso with overlap and graph lasso; p. 433-440.

Kaufman, L., Rousseeuw, P. Clustering by Means of Medoids. North-Holland; Amsterdam: 1987.

Kaufman, L., Rousseeuw, PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley; New York: 1990. MR1044997

Kim EY, Kim SY, Ashlock D, Nam D. MULTI-K: Accurate classification of microarray subtypes using ensemble *k*-means clustering. BMC Bioinform. 2009; 10:260.

Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]

Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, Mills KI, Gilkes AF, Hofmann WK, Basso G, Dell’Orto MC, Foà R, Chiaretti S, Vos JD, Rauhut S, Papenhausen PR, Hernández JM, Lumbreras E, Yeoh AE, Koay ES, Li R, Liu W, Williams PM, Wieczorek L, Haferlach T. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: The microarray innovations in LEukemia study prephase. Br J Haematol. 2008; 142:802–807. [PubMed: 18573112]

Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest. 2011; 121:2750. [PubMed: 21633166]

Lock EF, Dunson DB. Bayesian consensus clustering. Bioinformatics. 2013; 29:2610–2616. [PubMed: 23990412]

MacQueen, J. Proc Fifth Berkeley Sympos Math Statist and Probability (Berkeley, Calif, 1965/66). Univ California Press; Berkeley, CA: 1967. Some methods for classification and analysis of multivariate observations; p. 281-297.MR0214227

Maitra R, Ramler IP. Clustering in the presence of scatter. Biometrics. 2009; 65:341–352. MR2751457. [PubMed: 18537949]

McLachlan GJ, Bean R, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. 2002; 18:413–422. [PubMed: 11934740]

Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985; 50:159–179.

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009; 27:1160–1167. [PubMed: 19204204]

Parsons DW, Jones S, Zhang X, Lin JCH, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008; 321:1807–1812. [PubMed: 18772396]

Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics. 2006; 22:1988–1997. [PubMed: 16766561]

Ramasamy A, Mondry A, Holmes CC, Altman DG. Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med. 2008; 5:1320–1333.

Richardson S, Tseng GC, Sun W. Statistical methods in integrative genomics. Annu Rev Statist Appl. 2016; 3:181–209.

Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med. 2002; 346:1937–1947. [PubMed: 12075054]

Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LCG, Lannon WA, Grotzinger C, Del Rio M, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med. 2013; 19:619–625. [PubMed: 23584089]

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25:2906–2912. [PubMed: 19759197]

Shen K, Tseng GC. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. Bioinformatics. 2010; 26:1316–1323. [PubMed: 20410053]

Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. J Natl Cancer Inst. 2005; 97:866–867. [PubMed: 15956642]

Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst. 2003; 95:14–18. [PubMed: 12509396]

Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J Comput Graph Statist. 2013; 22:231–245. MR3173712.

Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, Kellam P. Consensus clustering and functional interpretation of gene-expression data. Genome Biol. 2004; 5:R94. [PubMed: 15535870]

Tibshirani R, Walther G. Cluster validation by prediction strength. J Comput Graph Statist. 2005; 14:511–528. MR2170199.

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; 63:411–423. MR1841503.

Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B, et al. Novel molecular subtypes of serous and endometrioid

ovarian cancer linked to clinical outcome. Clin Cancer Res. 2008; 14:5198–5208. [PubMed: 18698038]

Tseng GC. Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. Bioinformatics. 2007; 23:2247–2255. [PubMed: 17597097]

Tseng G, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res. 2012; 40:3785–3799. [PubMed: 22262733]

Tseng GC, Wong WH. Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. Biometrics. 2005; 61:10–16. MR2129196. [PubMed: 15737073]

Verhaak RG, Wouters BJ, Erpelinck CA, Abbas S, Beverloo HB, Lugthart S, Löwenberg B, Delwel R, Valk PJ. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. Haematologica. 2009; 94:131–134. [PubMed: 18838472]

Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA, IDH1, EGFR*, and *NF1*. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]

Wang SL, Liao LZ. Decomposition method with a variable parameter for a class of monotone variational inequality problems. J Optim Theory Appl. 2001; 109:415–429. MR1834183.

Witkos TM, Koscianska E, Krzyzosiak WJ. Practical aspects of microRNA target prediction. Curr Mol Med. 2011; 11:93–109. [PubMed: 21342132]

Witten DM, Tibshirani R. A framework for feature selection in clustering. J Amer Statist Assoc. 2010; 105:713–726. MR2724855.

Xie B, Pan W, Shen X. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. Electron J Stat. 2008; 2:168–212. MR2386092. [PubMed: 19920875]

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol. 2006; 68:49–67. MR2212574.

Zou H. The adaptive lasso and its oracle properties. J Amer Statist Assoc. 2006; 101:1418–1429. MR2279469.

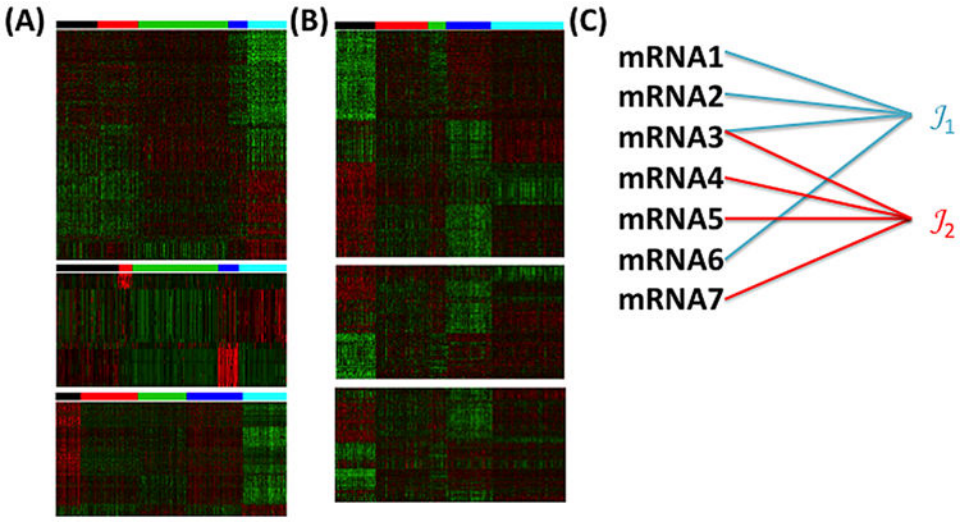

Fig. 1.

(A) Clustering of mRNA (upper heatmap) CNV (middle heatmap) and methylation (lower heatmap) profiles separately results in different five clusters of breast cancer subtypes (represented by color bars of five colors). (B) IS-Kmeans merges mRNA (upper heatmap) CNV (middle heatmap) and methylation (lower heatmap) and perform sample clustering. Inter-omics biological knowledge is also taken into account by overlapping group lasso. (C) An illustrating example of design of overlapping group lasso penalty term $\Omega(\mathbf{z})$ to incorporate prior knowledge of pathway information. Here,

$$\Omega(\mathbf{z}) = \sqrt{1+1+1/2+1}\sqrt{z_1^2+z_2^2+1/2\times z_3^2+z_6^2} + \sqrt{1/2+1+1+1}\times \sqrt{1/2\times z_3^2+z_4^2+z_5^2+z_7^2}.$$

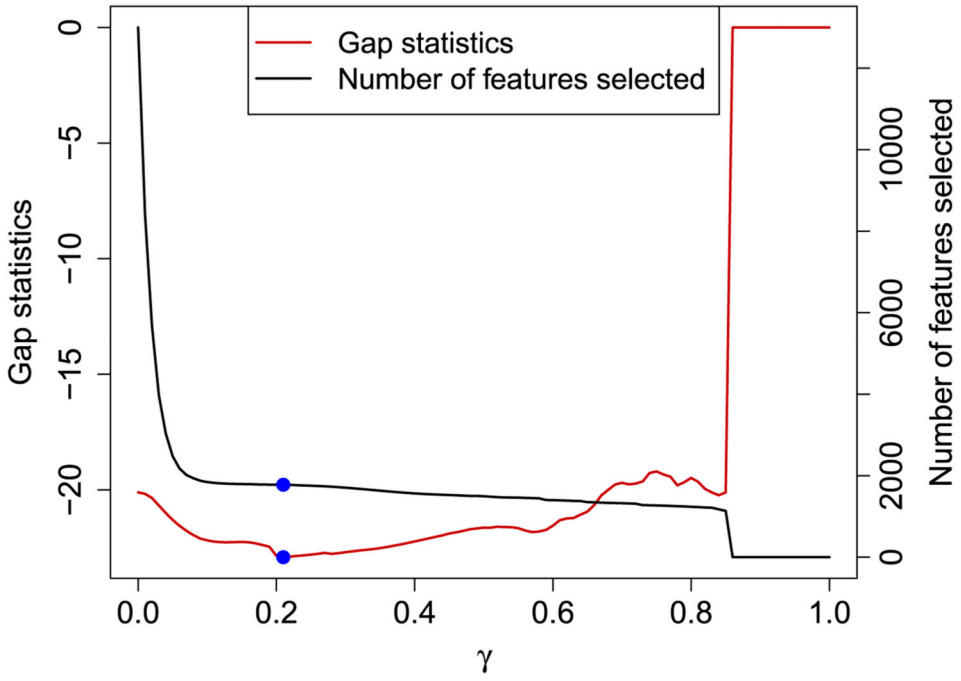

Fig. 2.

Selection oftuning parameter γ. This figure was from the simulated dataset in Section 4.1 with $\alpha = 0.5$. x-axis represents tuning parameter γ. Red curve and left y-axis denote the corresponding gap statistics. Black curve and right y-axis denote the corresponding number of selected features. The blue dots ($\gamma = 0.21$) represent where the gap statistics is minimized, and the corresponding number of selected feature is 1778.

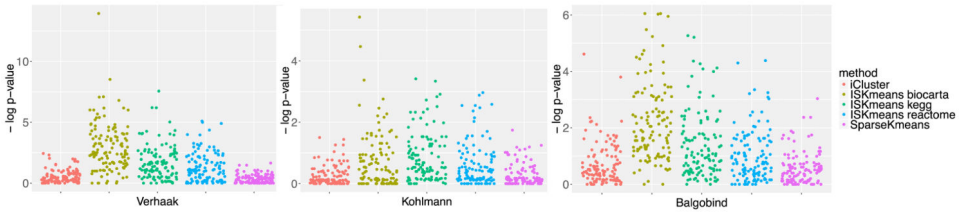

Fig. 3. Pathway enrichment analysis result for Leukemia BioCarta

Table 1

Comparison table of simulation with relative effect size f = 0.6. We simulated B = 100 times and calculated mean and standard deviation of each quantity. θ denotes the probability grouping information is correct for each feature inside groups. α is the tuning parameter balancing the emphasis between individual penalty and group penalty. For each method, we allow its own tuning parameter selection method to optimize its performance

θ	Method	α	ARI	Jaccard index	AUC	# features	Time (mins)
1	IS-*K* means	1	0.940 (0.239)	0.781 (0.202)	0.943 (0.138)	1465	0.44
		0.95	0.940 (0.239)	0.791 (0.204)	0.945 (0.136)	1483	0.52
		0.5	0.940 (0.239)	0.779 (0.202)	0.971 (0.084)	1420	0.56
		0.05	0.940 (0.239)	0.946 (0.214)	0.997 (0.012)	1723	0.67
0.2	IS-*K* means	1	0.940 (0.239)	0.781 (0.202)	0.943 (0.138)	1465	0.44
		0.95	0.940 (0.239)	0.783 (0.202)	0.943 (0.138)	1469	0.57
		0.5	0.940 (0.239)	0.602 (0.159)	0.943 (0.134)	1105	0.57
		0.05	0.940 (0.239)	0.467 (0.096)	0.888 (0.111)	2824	1.2
	iCluster		0.374 (0.323)	0.383 (0.274)		1239	26
	Sparse *K* means 1		0.312 (0.370)	0.105 (0.101)		896	0.12
	Sparse *K* means 2		0.361 (0.424)	0.204 (0.124)		2137	0.13

Table 2

Comparison of different methods using TCGA breast cancer (K = 5). G3 represents feature groups (gene symbol) where all three types of features are selected. Similarly, G2 represents feature groups (gene symbol) where only two types of features are selected; G1 represents feature groups (gene symbol) where only one type of feature is selected. We also compared the clustering result with PA M 50 subtype definition in terms of ARI

Method	ARI	nfeature	G1	G2	G3	Time
ISKmeans	0.379	2066	843	538	49	12.1 mins
SparseKmeans	0.332	2034	1466	284	0	6.85 mins
iCluster	0.272	2475	1725	375	0	3.91 hours

Table 3

Comparison of different methods using metabric breast cancer (K = 5). G2 represents feature groups (gene symbol) where all two types of features are selected; G1 represents feature groups (gene symbol) where only one type of feature is selected. The clustering result is compared with PA M 50 subtype definition in terms of ARI. Survival p-value obtained from the log rank test is given for the clustering assignment for each method

Method	ARI	nfeature	G1	G2	*p*-value	Time
ISKmeans	0.233	1882	1494	194	8.29×10^{-17}	38.4 mins
SparseKmeans	0.22	2004	2004	0	3.04×10^{-13}	34.3 mins
iCluster	0.0572	2471	2471	0	0.143	11.8 hours

Table 4

Comparison of different methods by ARI

		Verhaak		Kohlmann		Balgobind	
Method	**Pathway**	**# features**	**ARI**	**# features**	**ARI**	**# features**	**ARI**
	Biocarta	1009	0.932	1000	0.948	999	0.792
IS-*K* means	KEGG	1002	0.901	1013	0.948	990	0.792
	Reactome	993	0.932	994	0.948	1008	0.792
iCluster		982	0.733	1233	0.504	1020	0.214
sparse *K*-means		992	0.932	998	0.948	1014	0.792