Two-way Horizontal and Vertical Omics Integration for Disease Subtype Discovery

Zhiguang Huo[a*], Li Zhu[b], Tianzhou Ma[c], Hongcheng Liu[d], Song Han[e], Daiqing Liao[f], Jinying Zhao[g] and George Tseng[b*]

[a]Department of Biostatistics, University of Florida; [b]Department of Biostatistics, University of Pittsburgh; [c]Department of Epidemiology and Biostatistics, University of Maryland; [d]Department of Industrial and Systems Engineering, University of Florida; [e]Department of Surgery, University of Florida; [f]Department of Anatomy and Cell Biology, University of Florida; [g]Department of Epidemiology, University of Florida

ARTICLE HISTORY

Compiled May 1, 2019

ABSTRACT

Disease subtype discovery is an essential step in delivering personalized medicine. Disease subtyping via omics data has become a common approach for this purpose. With the advancement of technology and the lower price for generating omics data, multi-level and multi-cohort omics data are prevalent in the public domain, providing unprecedented opportunities to decrypt disease mechanisms. How to fully utilize multi-level/multi-cohort omics data and incorporate established biological knowledge toward disease subtyping remains a challenging problem. In this paper, we propose a meta-analytic integrative sparse Kmeans (MISKmeans) algorithm for integrating multi-cohort/multi-level omics data and prior biological knowledge. Compared with previous methods, MISKmeans shows better clustering accuracy and feature selection relevancy. An efficient R package, "MIS-Kmeans", calling C++ is freely available on GitHub (https://github.com/Caleb-Huo/MIS-Kmeans).

KEYWORDS

two-way horizontal and vertical omics integration; disease subtype discovery; ADMM; prior group information

1. Introduction

Disease subtyping is an essential step in delivering personalized medicine since different subtypes usually show strong clinical relevance and are in many cases responsive to different treatments (Abramson et al., 2015). Disease subtyping via omics data is prevalent in the literature, and representative studies include 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), ovarian cancer (Tothill et al., 2008), Parkinson's disease (Williams-Gray and Barker, 2017) and Alzheimer's disease (Bredesen, 2015). Using breast cancer as an example, the landmark paper by Perou et al. (2000) was among the first to identify five clinically meaningful subtypes (i.e., Luminal A, Luminal B,

[*]Correspondence: Z.H. zhuo@ufl.edu and T.G ctseng@pitt.edu

Her2-enriched, Basal-like and Normal-like) using gene expression profiles. Similar subtypes were identified by many independent studies afterwards (Ivshina et al., 2006; Loi et al., 2007; Sørlie et al., 2001; Van't Veer et al., 2002; Wang et al., 2005), and these subtyping results have been validated across studies with moderately satisfying consistency (Sørlie et al., 2003). However, different studies claim different intrinsic gene sets (i.e., a set of genes to define disease subtypes (Parker et al., 2009)). In addition, it has been pointed out that single cohort/single omics (e.g., transcriptome) analysis has limited sample size and suffers from reproducibility issues (Domany, 2014; Simon, 2005; Simon et al., 2003). Over the years, with the advancement of biotechnology (microarray and massively parallel sequencing), abundant data have been accumulated in public databases and repositories, including The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO) (Edgar et al., 2002) and Sequence Read Archive (SRA) (Kodama et al., 2011). These datasets provide unprecedented opportunities to decrypt disease mechanisms via integrating multiple cohorts or multiple-level omics data types (Tseng et al., 2012) (see Figure 1(A) for illustration of multiple cohorts or multiple-level omics data layout). Properly integrating these complex datasets will strengthen statistical power toward biological findings (Richardson et al., 2016). On the other hand, a tremendous amount of biological knowledge has been established through these datasets (e.g., gene pathway information, miRNA targeting gene databases, and the cis-acting regulatory mechanism of a certain gene) (Fan and Kurgan, 2015; Witkos et al., 2011). An example of group structure for prior biological knowledge is shown in Figure 1(B) and 1(C). In Figure 1(B), a group represents the potential cis regulatory relationship between mRNA, methylation and copy number variation (CNV) of the same gene symbol. In Figure 1(C), a group is a biological pathway including a collection of functionally related genes, which can potentially overlap with other pathways. Proper use of this prior knowledge can greatly facilitate modeling integrative analysis (Huo et al., 2017).

In the literature, various types of omics data integration approaches have been proposed. Tseng et al. (2012) categorized omics data integration into two major types: horizontal omics meta-analysis and vertical omics integrative analysis. On one hand, horizontal omics meta-analysis aims to combine multiple studies of the same omics data type (e.g., gene expression data from multiple studies as illustrated in dashed rectangle I in Figure 1(A) horizontally). This approach has been widely adopted to increase statistical power and reproducibility for differential expression analysis (Choi et al., 2003; Ramasamy et al., 2008), pathway enrichment analysis (Shen and Tseng, 2010), network analysis (Danaher et al., 2014; Zhu et al., 2016), clustering analysis (Huo et al., 2016), and dimension reduction (Kim et al., 2017). On the other hand, vertical omics integrative analysis aims to integrate multi-levels of omics data from the same patient cohort (Richardson et al., 2016) (e.g., genome-wide profiling of gene expression, DNA copy number variation, methylation of the same study as illustrated in dashed rectangle II in Figure 1(A) vertically). The idea of directly combining multi-levels of omics data has been extended into association analysis (Bottolo et al., 2013), regression (Kim et al., 2012; Wang et al., 2012), and clustering. Others seek to integrate multi-levels of omics data via incorporating prior knowledge (Huo et al., 2017; Quintana and Conti, 2013; Stingo et al., 2011) as illustrated in Figure 1(B) and 1(C). Readers can refer to Huang et al. (2017); Li et al. (2016); Richardson et al. (2016) for comprehensive reviews of existing omics integration methods.

In the realm of integrative clustering, several methods have been proposed. Lock and Dunson (2013) performed Bayesian consensus clustering by fitting a finite Dirichlet mixture model, which allowed both common and omics-type-specific clustering, but this model did not consider feature selection and is thus not suitable for high-dimensional omics data. Shen et al. (2012) fitted an integrative latent variable factor model (iCluster), but the method did not incorporate prior biological knowledge and required extensive computing with large matrix operations. Wang et al. (2014) proposed a network fusion approach to aggregate the sample similarity in each omics type, but this approach did not perform feature selection thus it was lack of biological interpretation about which genes would contribute to the subtype results. Huo et al. (2017) fitted an integrative sparse Kmeans model by incorporating prior biological knowledge by overlapping group lasso. However, to the best of our knowledge, no one has performed a two-way omics integration to fully utilize multi-cohort and multi-omics information. Hence, in this paper we propose a meta-analytic multi-omics data integration framework (MISKmeans) to perform disease subtype discovery (sample clustering), extending from previous work on the meta-analytic sparse Kmeans algorithm (MetaSparseKmeans) (Huo et al., 2016) and the integrative sparse Kmeans algorithm (ISKmeans) (Huo et al., 2017). The major novelty is that this algorithm is the first to simultaneously accommodate both horizontal omics meta-analysis and vertical omics integrative analysis, closing the research gap between horizontal and vertical omics data integration. Prior biological information can also be incorporated to guide feature selection using overlapping group lasso (Jacob et al., 2009). The complex optimization problem is efficiently solved using the alternating direction method of multiplier (ADMM) (Boyd et al., 2011). Previously, MetaSparseKmeans (Huo et al., 2016) required the computational complexity of order $(K!)^{S-1}$ to match the clustering patterns. This is improved in this paper by efficient memory design so that the complexity reduces to the order of $S(S-1)K!/2$. Such improvement makes it feasible to apply MISKmeans in moderate scale subtype discovery analysis. In our simulation and real data applications, we not only show striking clinical differences between the resulting subtypes, but also demonstrate better performance than the previous MetaSparseKmeans algorithm (Huo et al., 2016) and the ISKmeans algorithm (Huo et al., 2017) in terms of clustering accuracy and feature selection relevance. These appealing results are expected since the proposed algorithm simultaneously integrates multi-cohort, multi-omics and prior biological information and thus generates the most reliable and comprehensive result.

2. Method

2.1. Kmeans algorithm and its derivatives for omics data integration

2.1.1. Kmeans algorithm.

Consider X_{jl} the gene expression level of gene j and sample l. Note that we use gene expression as an example, and it can be replaced with other types of omics data. Denote within cluster sum of square of gene j as $WCSS_j(C) = \sum_{k=1}^{K} \sum_{l \in C_k} (X_{jl} - \bar{X}_{jC_k})^2$, where $\bar{X}_{jC_k}$ is the center for cluster k and gene j; K is the number of clusters; $C = (C_1, \ldots, C_k, \ldots, C_K)$ is the clustering result, with C_k indicating a collection of subjects of cluster k. The Kmeans algorithm (Hartigan and Wong, 1979) obtains the clustering results by minimizing the WCSS:

$$\min_C \sum_{j=1}^{p} WCSS_j(C)$$

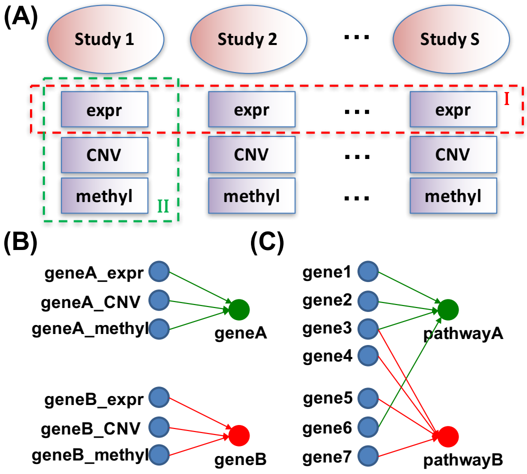

Figure 1. (A) Illustration of multiOmics meta-analytic data integration. (B) A group contains gene expression, CNV and methylation of the same gene symbol. (C) A group is a pathway, which is a collection of genes (e.g., cell cycle pathway).

2.1.2. Sparse Kmeans algorithm.

Since the total cluster sum of square ($TSS_j = \sum_{l \in C}(X_{jl} - \bar{X}_{jC})^2$) can be decomposed as between cluster sum of square (BCSS) plus within cluster sum of square (WCSS), minimizing the WCSS is equivalent to maximizing the BCSS as $BCSS_j(C) = TSS_j - WCSS_j(C)$. A sparse Kmeans algorithm was proposed (Witten and Tibshirani, 2010) by adding gene specific weight to BCSS and imposing lasso regularization on gene specific weight:

$$\min_{C,\mathbf{z}} - \sum_{j=1}^{p} z_j BCSS_j(C) + \gamma \|\mathbf{z}\|_1$$
$$\text{subject to } \|\mathbf{z}\|_2 \le 1, z_j \ge 0, \forall j, \tag{1}$$

where z_j denotes the weight for gene j, $\|\mathbf{z}\|_1$ and $\|\mathbf{z}\|_2$ represent the l_1 norm and l_2 norm of the weight vector $\mathbf{z} = (z_1, \ldots, z_p)$ and γ is a tuning parameter for the l_1 norm penalty. By iteratively optimizing clustering assignment C and weight vector $\mathbf{z}$, one can obtain both clustering result C and a sparse solution of feature selection (non-zero element of $\mathbf{z}$).

2.1.3. Meta-analytic sparse Kmeans algorithm.

Huo et al. (2016) extended Equation 1 toward a meta-analytic framework of sparse clustering (MetaSparseKmeans) by combining multiple cohorts using standardized BCSS (standardized by TSS), by which potential batch effect can be circumvented, and different studies are on a comparable scale. The standardized BCSS is defined as $R_j(C) = BCSS_j(C)/TSS_j$ (ranges between 0 to 1), which measures the separating ability of each gene feature. The algorithm is represented as:

$$\min_{C^{(s)},\mathbf{z},M} -\sum_{j=1}^{p} z_j \times \left[\frac{1}{S}\sum_{s=1}^{S} R_j^{(s)}(C^{(s)}) + \lambda \times f_j(M)\right] + \gamma\|\mathbf{z}\|_1$$
$$\text{subject to } \|\mathbf{z}\|_2 \leq 1, z_j \geq 0, \tag{2}$$

where $R_j^{(s)}(C^{(s)})$ denotes the separating ability of gene j and study s, with clustering assignment $C^{(s)}$, with $C^{(s)} = \{C_1^{(s)}, \ldots, C_K^{(s)}\}$, and $C_k^{(s)}$ $(1 \leq k \leq K)$ is a collection of samples within the k^{th} subtypes (clusters) of cohort s. Note K is assumed to be equal across all cohorts since we expect the same number of subtypes for a common disease. In order to guarantee that the resulting subtype patterns are consistent across studies, a pattern matching award function $f_j(M)$ is introduced for feature j, where M is a cluster matching rule across S studies. For instance, when $S = 2$ and $K = 3$, denote $M_1 = (C_1^{(1)} - C_1^{(2)}, C_2^{(1)} - C_3^{(2)}, C_3^{(1)} - C_2^{(2)})$ as a possible matching rule, where first cluster in study 1 matches to first cluster in study 2, second cluster in study 1 matches to third cluster in study 2, and third cluster in study 1 matches to second cluster in study 2. Similarly, $M_2 = (C_1^{(1)} - C_1^{(2)}, C_2^{(1)} - C_2^{(2)}, C_3^{(1)} - C_3^{(2)})$ is another possible matching rule. **Figure S1** shows a concrete example of the pattern matching based on M_1 and M_2. In this example, M_2 is favored over M_1, as it provides consistent pattern matching across subtypes (also see detailed explanation in **Supplementary Section I.1**). For each pair of study s and s', the pattern award function, denoted as $h_j^{(s,s')}(M)$, is defined as the multi-class correlation (MCC) (Huo et al., 2016; Lu et al., 2009), and the details are described in **Supplementary Section I.1**. $f_j(M) = \frac{2}{S(S-1)}\sum_{s<s'} h_j^{(s,s')}(M)$ is the pattern matching reward function aggregating all pairs of studies s, s'. By definition $f_j(M)$ ranges from 0 to 1, with the larger value representing more consistent clustering pattern across all studies, the separating ability $\frac{1}{S}\sum_{s=1}^{S} R_j^{(s)}$ and the matching ability $f_j(M)$ are comparable since both of them range from 0 to 1. By minimizing Equation 2, we simultaneously obtain feature selection, a clustering pattern in each cohort and subtype patterns match rules, which defines final subtypes across all studies.

2.1.4. Integrative sparse Kmeans algorithm.

Huo et al. (2017) extended Equation 1 to group structured sparse Kmeans (ISKmeans). Under this scenario, we consider p as the total number of features to be combined, including all levels of omics datasets. We further impose the overlapping group lasso penalty term $\Omega(\mathbf{z})$ in the objective function, which encourages features from the same group to be selected together. Such overlapping group structure is shown and explained in Figure 1(B) and 1(C), in which a group contains multiple levels of omics features of the same gene symbol or a collection of genes inside a pathway. The objective function of ISKmeans is shown below:

$$\min_{C,\mathbf{z}} -\sum_{j=1}^{p} z_j R_j(C) + \gamma\alpha\|\mathbf{z}\|_1 + \gamma(1-\alpha)\Omega(\mathbf{z})$$
$$\text{subject to } \|\mathbf{z}\|_2 \leq 1, z_j \geq 0, \forall j, \tag{3}$$

where γ is still the tuning parameter controlling the number of nonzero features, and $\alpha \in [0,1]$ is a term balancing between individual feature penalty and group feature penalty. If $\alpha = 1$, only the individual feature penalty is imposed, and if $\alpha = 0$, only the group feature penalty is imposed. The overlapping group lasso penalty is defined as $\Omega(\mathbf{z}) = \sum_{1\leq g\leq G_0} w_g \|\mathbf{m}_g \circ \mathbf{z}\|$, where G_0 is the total number of prior groups (potentially overlapping), $w_g \in \mathbb{R}$ is the group level weight coefficient for group g, $\mathbf{m}_g \in \mathbb{R}^p$ is the feature level design vector of the group g and $\circ$ denotes Hadamard product. For example, there are three features $\{1,2,3\}$ and two groups J_1 and J_2, where $J_1 = \{1,2\}$ and $J_2 = \{2,3\}$ (J_1 and J_2 overlap with feature $\{2\}$). Under this scenario, the overlapping group lasso penalty $\Omega(\mathbf{z}) = w_1\sqrt{m_1^2 z_1^2 + m_2^2 z_2^2} + w_2\sqrt{m_2^2 z_2^2 + m_3^2 z_3^2}$. w_g and $\mathbf{m}_g$ have to be designed carefully, otherwise bias will be introduced toward feature selection. Below is an illustrating example to reflect the potential bias in the coefficient design. Firstly, if w_2 is larger than w_1, J_2 will be penalized more than J_1, and thus features in group J_2 are less likely to be selected. Secondly, when fixing $w_1 = w_2$, if we assign $m_1 = m_2 = m_3$, feature $\{2\}$ will be over penalized since m_2 involves in both J_1 and J_2. It is a challenging question how to design these coefficients such that the feature selection is unbiased under the overlapping group case. Huo et al. (2017) has proposed a design of overlapping group lasso penalty for w_g and $\mathbf{m}_g$, which satisfies the “unbiased feature selection principle”. This unbiased design is also adopted in MISKmeans and details will be introduced in the **Supplementary Section I.2**.

2.2. Objective function of MISKmeans

We propose the MISKmeans algorithm objective function by further extending Equation 2 and Equation 3.

$$\begin{aligned} &\min_{C^{(s)},\mathbf{z},M} -\sum_{j=1}^{p} z_j \times \left[\frac{1}{S}\sum_{s=1}^{S} R_j^{(s)} + \lambda \times f_j(M)\right] + P(\mathbf{z}) \\ &\text{s.t. } \|\mathbf{z}\|_2 \leq 1, z_j \geq 0, P(\mathbf{z}) = \gamma\alpha\|\mathbf{z}\|_1 + \gamma(1-\alpha)\Omega(\mathbf{z}) \end{aligned} \tag{4}$$

The objective function of MISKmeans in Equation 4 generates a common set of intrinsic features from the non-zero estimated weight z_j for all S cohorts. The second term $R_j^{(s)}$ measures standardized BCSS of study s and feature j, and minimizing $-R_j^{(s)}$ yields good sample clustering separation in each cohort. The third term $f_j(M)$ guarantees the clustering patterns are consistent across cohorts, which leads to a definition of common disease subtype. The last term $P(\mathbf{z})$ is composed of two terms: a l_1 norm penalty term $\|\mathbf{z}\|_1$ generating sparsity on feature weights to facilitate feature selection and an overlapping group lasso term $\Omega(\mathbf{z})$ encouraging features belonging to the same group to be selected together. Such a penalty design guarantees that a small set of informative features will be selected. The overlapping group lasso penalty is $\Omega(\mathbf{z}) = \sum_{1\leq g\leq G_0} w_g\|\mathbf{m}_g \circ \mathbf{z}\|$ (the same as in the previous subsection), and the “unbiased feature selection principle” are similarly adopted for the coefficient design of w_g and $\mathbf{m}_g$. Details about the definition of “unbiased feature selection principle” and the coefficient design for $\Omega(\mathbf{z})$ are given in the **Supplementary Section I.2**. Further, MetaSparseKmeans (Equation 2) and ISKmeans (Equation 3) are special cases of MISKmeans (Equation 4). MISKmeans (Equation 4) reduces to MetaSparseKmeans when $\alpha = 1$ and reduces to ISKmeans when $S = 1$.

2.3. Tuning parameter selection

K, λ, γ and α need to be estimated in Equation 4.

2.3.1. Selection of K

K is assumed to be the same for all cohorts (same number of subtypes in different studies). In the literature, there are two types of integrative clustering algorithms: (1) assuming equal number of clusters for different layers of omics data. In line with this assumption, there are iCluster (Shen et al., 2012), ISKmeans (Huo et al., 2017), and similarity network fusion (Wang et al., 2014); (2) assuming unequal number of clusters for different layers of omics data. In line with this assumption, there is Bayesian concensus clustering (Lock and Dunson, 2013). The proposed MISkmeans extends from ISKmeans, which assumes common number of clusters. The issue of estimating K has been widely discussed in the literature and has been well-recognized as a difficult and data-dependent problem (Kaufman and Rousseeuw, 2009; Milligan and Cooper, 1985). Here, we suggest to use gap statistics (Tibshirani et al., 2001) in individual studies and make a joint decision. For example, one can perform gap statistics using the sparse Kmeans as the clustering algorithm to determine K in each individual studies. And a consensus K can be determined by majority voting from individual studies. However, in the case where no consensus K can be determined, we suggest to use domain knowledge to determine K, or try multiple Ks and determine the best choice via other external biological knowledge.

2.3.2. Selection of λ

λ balances the separation ability and clustering pattern matching ability. Huo et al. (2016) has pointed out that λ is not sensitive to the performance of the clustering results and has suggested the choice of $\lambda = 0.5$. In this paper, we further proposed a post selection algorithm to determine λ such that $\sum_j R_j^{(s)}(C^{(s)}) = \lambda \sum_j f_j(M)$ given one preset MISKmeans result. In Section 3.1.2, we have performed simulations with various signal strengths and prior group information accuracy. This sensitivity analysis shows that the choice of λ is not very sensitive for the performance of clustering accuracy. Fixing $\lambda = 0.5$ or choosing λ by the proposed selection criteria usually perform well. As a result, we apply $\lambda = 0.5$ throughout the paper for computational convenience unless otherwise indicated.

2.3.3. Selection of α

α balances between individual feature penalty and group penalty in the sparse group lasso (Simon et al., 2013). According to the Equation 4, $\alpha = 1$ means we only emphasize on individual feature penalty and ignore overlapping group penalty, in which the MISkmeans is equivalent to MetaSparseKmeans. $\alpha = 0$ means we only emphasize overlapping group penalty and ignore individual feature penalty. Simon et al. (2013) argued that there is no theoretically optimal selection for α since selection of α relates to multiple factors such as accuracy of prior group information and sparsity within groups. Choice of α depends on whether the grouping information is correct. In this paper, we further propose a post selection algorithm to determine α such that $\alpha\|\mathbf{z}\|_1 = (1-\alpha)\Omega(\mathbf{z})$ given one preset MISKmeans result. The sensitivity analysis in Section 3.1.2 shows that smaller α is preferred given correct grouping information, and larger α is preferred given partially correct grouping information. We further find that fixing $\alpha = 0.5$ or choosing the proposed selection criteria usually perform well regardless of the grouping information is correct or partially correct. As a result, we apply $\alpha = 0.5$ throughout the paper for computational convenience unless otherwise indicated.

2.3.4. Selection of γ

γ is the the penalty coefficient to control number of selected features. When γ is large, we place large penalty on the objective function and end up with less selected features. When γ is small, we place small penalty and will include more features. We follow and extend the gap statistic procedure (Huo et al., 2016; Tibshirani et al., 2001) to estimate γ:

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

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

$$\text{Gap}(\gamma) = O(\gamma) - \frac{1}{B}\sum_{b=1}^{B} O_b(\gamma), \tag{5}$$

where $O(\gamma) = -\sum_{j=1}^{J} z_j^* R_j(C^*)$ is from observed data, where $\mathbf{z}^*, C^*$ are the minimizer of the objective function in Equation 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 5 is minimized.

In practice, calculating gap statistics from a chain of γ can be performed efficiently by adopting 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. An example of gap statistics result for our simulation is shown in **Figure S2**, where the gap statistics successfully uncover the underlying 1,800 subtype predictive genes.

2.4. Optimization

Three parameters $C^{(s)}, \mathbf{z}, M$ in objective function of Equation 4 are updated iteratively until convergence. Below is the detailed optimization procedure:

(1) Estimate $\mathbf{z}_s$ in the s^{th} cohort using sparse Kmeans algorithm. Set initial value of $\mathbf{z} = \frac{1}{S}\sum_{i=1}^{S}\mathbf{z}_s$.
(2) Fix $\mathbf{z}$; update $C^{(s)}$ in study $s (1 \leq s \leq S)$ by weighted Kmeans.
(3) Fix $\mathbf{z}$ and $C^{(s)}$; update M by exhaustive search.
(4) Fix $C^{(s)}$ and M; update $\mathbf{z}$.
(5) Iterate Step 2 through Step 4 until convergence.

In Step 2, it is well acknowledged that most Kmeans derived algorithms suffer from the local optimum problem. A commonly adopted approach is to perform multiple initializations and select the clustering result with the best objective score. Our R package sets 20 initializations as the default for this Kmeans step, but the user can further tune this parameter to avoid the local minimum problem as much as possible. For Step 3, Huo et al. (2016) claimed total searching configuration as $(K!)^{S-1}$, which is extremely hard when K and S are large. In our new implementation, we bring down the computational complexity to $S(S-1)K!/2$ by caching all pairwise MCC $h_j^{(s,s')}(M)$. Since all pairwise MCC are cached, the complexity of the exhaustive search itself is ignorable compared to the caching step, whose complexity is $S(S-1)K!/2$. Such improvement makes it feasible to apply MISKmeans in moderate scale analysis. Step 4 is a challenging convex optimization problem since sparse overlapping group lasso penalty is involved. We adopt the alternating direction method of multipliers (ADMM) to efficiently solve this problem (Boyd et al., 2011; Huo et al., 2017).

3. Results

In order to evaluate the two-way integration approach and compare to the one-way integration (either horizontal or vertical), we compared the performance of MISKmeans, ISKmeans and MetaSparseKmeans in simulation studies. Comparisons with other one-way integrative clustering method have been discussed previously (Huo et al., 2017; Wang et al., 2014). For example, ISKmeans was shown to outperform iCluster in Huo et al. (2017) and the comparison will not be repeated here. Our proposed algorithm is further evaluated in two real datasets, including multi-cohort multi-omics breast cancer data using multiple levels of omics features of the same gene symbol as group information and multi-cohort leukemia data using external pathway database as group information. The datasets description is shown in **Table S1**. Note that we have set 20 initializations for all Kmeans related procedures to try to avoid local optimum problems.

3.1. Simulation study

3.1.1. Main simulation result.

To compare the performance of MISKmeans, ISKmeans and MetaSparseKmeans, we designed simulations with details described in **Supplementary Section II**. To be brief, we simulated three multiOmics studies, with each study containing two omics types (gene expression data and DNA methylation data). Each omics type contained (a) 900 subtype predictive genes – genes that define the underlying subtypes; (b) 2,400 confounder impacted genes (e.g., gender, race, other demographic factors or disease stage, etc.), which added heterogeneity to each study to complicate disease subtype discovery; (c) 5,000 non-informative genes – random noise. There were totally 16,600 features and more than 100 subjects for each study. The simulation also imposed correct group structure between the subtype predictive genes from two omics types, which were prior knowledge fed to MISKmeans and ISKmeans. Further, we used a parameter f to denote the subtype separation ability, with large f indicating stronger separation ability. For a fair comparison, we also implemented the MetaSparseKmeans algorithm with the improved pattern matching complexity (i.e., $S(S-1)K!/2$).

In order to determine number of clusters K from the data, we selected 500 features of largest variance to perform gap statistics using the sparse Kmeans as clustering algorithm, in each study and each omics type. The resulting gap statistics (**Figure S3**) implies $K = 3$ is optimal, which is consensus among majority of studies/omics types.

Table 1. Comparison table of simulation for MISKmeans, ISKmeans and MetaSparseKmeans (MSKM), with relative effect size $f = 0.6$, 0.4 and 0.3. We simulated $B = 100$ times and calculated mean and (standard deviation) of each quantity. For each method, we selected the tuning parameter such that the number of selection features were closest to the underlying truth.

f	method	ARI	Jaccard index	AUC	time (min)
0.6	MISKmeans	1 (0)	1 (0)	1 (0)	1.52
	ISKmeans	1 (0)	0.92 (0.02)	0.97 (0.01)	1.32
	MSKM	0.85 (0.24)	0.8 (0.22)	0.92 (0.11)	1.28
0.4	MISKmeans	0.99 (0.01)	1 (0)	1 (0)	1.84
	ISKmeans	0.98 (0.02)	0.82 (0.03)	0.92 (0.02)	1.33
	MSKM	0.74 (0.25)	0.67 (0.23)	0.89 (0.13)	1.54
0.3	MISKmeans	0.94 (0.03)	1 (0)	1 (0)	2.42
	ISKmeans	0.63 (0.18)	0.55 (0.11)	0.78 (0.05)	1.4
	MSKM	0.36 (0.19)	0.4 (0.16)	0.8 (0.11)	1.64

To benchmark the performance, we used ARI (Hubert and Arabie, 1985) and Jaccard index (Jaccard, 1901) to evaluate the clustering and feature selection performance. ARI calculates the consistency of the clustering result with the underlying true clustering in simulation (the range is -1 to 1, and 1 represents exactly the same partition as the underlying truth). The Jaccard index compares 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 (the range is 0 to 1, and 1 represents identical feature sets compared to the underlying truth). Note that MISKmeans integrated 3 multiOmics studies simultaneously, ISKmeans integrated each of the multiOmics studies individually, then averaged the results from all studies, and the MetaSparseKmeans integrated three studies for each omics type respectively and then averaged the results from all omics types. To eliminate the confounding effect of tuning parameter selection, we pre-selected a wide range of tuning parameters for all methods and chose the best tuning parameter such that the number of selected features were closest to the underlying truth for each method. The resulting ARI and Jaccard index are given in Table 1. Clearly, MISKmeans outperforms ISKmeans and MetaSparseKmeans, especially when the signal level f is small. We further compared feature selection in terms of area under the curve (AUC) of ROC curve, which avoided the issue of tuning parameter selection. Here, the sensitivity and specificity of the ROC curve were calculated by comparing the selected features to the underlying subtype predictive genes. Remarkably, MISKmeans beats ISKmeans MetaSparseKmeans with perfect AUCs. In addition, all three methods were very fast ($1 \sim 2$ minitus) for integrating high-dimensional multiOmics data. Note that the time was calculated by summing up all individual study evaluation time for ISKmeans or individual omics type evaluation time for MetaSparseKmeans.

3.1.2. Sensitivity analysis of λ and α.

To evaluate the impact of different choices of λ and α; the performance of the default choice of $\lambda = 1/2$, $\alpha = 1/2$; and the performance of the λ and α by the proposed selection criteria, we performed simulations in the following 4 special settings. These special simulation settings inherit the general procedure of the main simulation (e.g., $S = 3$, $T = 2$), but vary in the following aspects:

(a) $f_1 = 0.6$, $f_2 = 0.6$, $f_3 = 0.4$, and $\theta = 1$.

(b) $f_1 = 0.6$, $f_2 = 0.6$, $f_3 = 0.4$, and $\theta = 0.6$.
(c) $f_1 = 0.4$, $f_2 = 0.4$, $f_3 = 0.2$, and $\theta = 1$.
(d) $f_1 = 0.4$, $f_2 = 0.4$, $f_3 = 0.2$, and $\theta = 0.6$.

The candidate $\lambda = 0.1, 0.2, 0.5, 1, 2$ and the λ^* by the selection criteria proposed in Section 2.3.2; The candidate $\alpha = 0.05, 0.5, 0.95$, and the α^* by the selection criteria proposed in Section 2.3.3 were jointly evaluated. These simulations were repeated for $B = 50$ times. The performance is shown in **Table S2** and **S3**. We have the following observations from these tables: (a), when studies have strong to moderate signals ($f = 0.6 \sim 0.4$) and the grouping information is correct ($\theta = 1$), the clustering accuracy is very high regardless different choices of α and λ, and the feature selection accuracy is better given smaller α (e.g., $\alpha = 0.05$ or $\alpha = 0.5$); (b), when studies have strong to moderate signals ($f = 0.6 \sim 0.4$) and the grouping information is partially correct ($\theta = 0.6$), the clustering accuracy is very high regardless different choices of α and λ, and the feature selection accuracy is better given larger α (e.g., $\alpha = 0.5$ or $\alpha = 0.95$); (c), when studies have moderate to weak signals ($f = 0.4 \sim 0.2$) and the grouping information is correct ($\theta = 1$), the clustering accuracy are similar for different choices of α and λ, and the feature selection accuracy is better given smaller α (e.g., $\alpha = 0.05$ or $\alpha = 0.5$); (d), when studies have moderate to weak signals ($f = 0.4 \sim 0.2$) and the grouping information is partially correct ($\theta = 0.6$), the clustering accuracy are similar for different choices of α and λ, and the feature selection accuracy is better given larger α (e.g., $\alpha = 0.5$ or $\alpha = 0.95$). Regardless of various simulation settings, the default tuning parameter $\alpha = 0.5$ and $\lambda = 0.5$, as well as the tuning parameter by the proposed selection criteria (marked by $*$) generally perform very good.

Collectively, the choice of λ is not very sensitive for the performance of clustering accuracy. Fixing $\lambda = 0.5$ or choosing λ by the proposed selection criteria usually perform well. Choice of α depends on whether the grouping information if correct. Smaller α is preferred given correct grouping information, and larger α is preferred given partially correct grouping information. Fixing $\alpha = 0.5$ or choosing α by the proposed selection criteria usually perform well for both scenarios. For computational convenience, we will fix $\lambda = 0.5$ and $\alpha = 0.5$ for all other evaluations.

3.2. Multi-cohort multi-omics breast cancer example

In this breast cancer example, we combined TCGA data (Weinstein et al., 2013) and METABRIC data (Curtis et al., 2012), with gene expression and copy number variation in both of them. The gene expression data from array and RNA-seq were transformed in log-scale, and the CNV data were measured by segment mean values, which are equal to log2(copy-number/2). For METABRIC, we adopted the same protocol (Curtis et al., 2012) to split the cohort into a discovery set and validation set. Detailed descriptions regarding platforms and the number of features for each level of omics data are available in **Table S1(a)**. Detailed preprocessing procedures are given in the **Supplementary Section III**. We applied MISKmeans on the multi-cohort (TCGA + METABRIC discovery set) multi-omics data and compared with ISKmeans (applied to METABRIC discovery only) and MetaSparseKmeans (ignore the grouping). In order to determine number of clusters K from the data, we selected 500 features of largest variance to perform gap statistics using the sparse Kmeans as clustering algorithm, in each study and each omics type. However, in **Figure S4**, we didn't have a consensus K that is supported in all cases. Instead, we chose $K = 5$ since it is well known that there are five subtypes of breast cancer by PAM50 definition

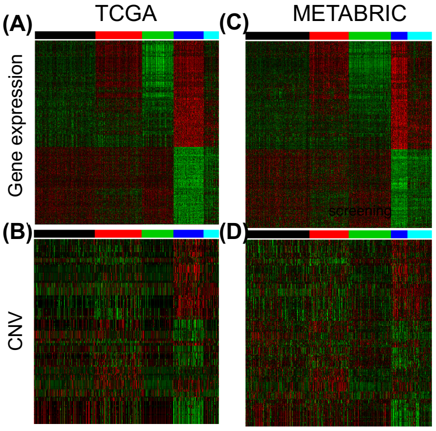

Figure 2. Multi-cohort multi-omics clustering results using breast cancer data. The color bar on top of heatmap denotes distinct subtypes. (A) TCGA gene expression. (B) TCGA CNV. (C) METABRIC gene expression. (D) METABRIC CNV.

(Parker et al., 2009). For a fair comparison, we chose the tuning parameter such that each method ends up with about 1,000 features.

The resulting multi-omics profiles of MISKmeans are depicted in Figure 2. We obtained 5 distinct subtypes from TCGA and METABRIC discovery multi-omics data respectively, and the clustering patterns were consistent across the two cohorts. To benchmark the performance, we evaluated survival differences among the subtypes obtained from MISKmeans in the METABRIC discovery cohort. The resulting p-value for the survival difference among five subtypes from likelihood ratio test after adjusting for treatment (chemotherapy) is 1.35×10^{-8}, indicating that the resulting subtypes are have distinct survival difference.

We further compared MISKmeans with ISKmeans (only on METABRIC discovery set) and MetaSparseKmeans in terms of feature selection and clustering accuracy. For feature selection, since we imposed the cis-regulatory relationship between mRNA and CNV as prior knowledge, we wanted to confirm whether such pairs of features were more frequently selected by MISKmeans. Therefore, we investigated two categories of feature groups defined in Huo et al. (2017): G1 and G2. G2 represents a feature group where both mRNA and CNV of the same gene symbol are selected, while G1 represents a feature group where either mRNA or CNV is selected. The comparison results of MISKmeans, ISKmeans and MetaSparseKmeans in terms of feature selection are shown in Table 2. Clearly, MISKmeans and ISKmeans obtained more G2 features than MetaSparseKmeans. This is biologically more interpretable but not surprising since MISKmeans and ISKmeans utilized multi-omics regulatory information, and features of the same group are expected to be selected together. Remarkably, MISKmeans selects more G2 features than ISKmeans, indicating the potential enhancement of feature selection with multi-omics meta-analytic integration.

For clustering accuracy, we do not know the underlying truth, but survival separation is a clinical relevant benchmark to compare different methods. Table 2 shows that the p-value of survival difference (adjusted for chemotherapy) for the clustering results defined by MISKmeans is slightly more significant than the other methods. We also compared the clustering result with the PAM50 subtype definition in terms of

Table 2. Comparing Performance on BRCA in terms of clustering consistency (in METABRIC discovery set) with PAM50 (measured by ARI), number of cis-regulatory groups (G2) and survival difference of 5 groups. Survival p-values are measured in -log_{10} scale. Time is measured in minutes.

method	nfeature	G1	G2	time	Silhouette	ARI	Survival
MISKmeans	992	334	329	30.8	0.11	0.34	7.87
ISKmeans	911	805	53	6.11	0.09	0.25	7.77
MSKM	987	977	5	28.3	0.11	0.34	7.11
PAM50							6.06

ARI in METABRIC discovery cohort. MISKmeans and MetaSparseKmeans achieve higher ARI compared to ISKmeans using PAM50 as benchmark. The five-by-five confusion table of MISKmeans clustering result and PAM50 subtypes is in **Table S4**. Note that the ARI for PAM50 and the three integrative methods is not very high ($0.25 \sim 0.34$), which may be due to the fact that PAM50 was defined by gene expression only, but in our case multi-omics data are integrated. Since the p-value for survival difference for MISKmeans is more significant than that of PAM50, the subtypes defined by MISKmeans may be clinically more meaningful than that of PAM50. This may indicate certain machine learning methods could achieve better breast cancer subtype definitions. In addition, we used Silhouette scores (Rousseeuw, 1987) to assess the coherence of the resulting clustering. We found MISKmeans and MetaSparseKmeans achieved higher average Silhouette scores than ISKmeans, meaning that combining multiple cohorts indeed enhanced clustering result. Computing time of all three methods are within around 30 minutes.

Further, we applied weighted Kmeans in the METABRIC validation cohort, using the selected genes from each method respectively. The resulting survival comparison p-values, ARI and Silhouette scores are shown in **Table S5**. Again, we observe MISKmeans and MetaSparseKmeans outperform ISKmeans.

3.3. Multi-cohort leukemia transcriptomic datasets using pathway database as prior knowledge

In the previous multi-cohort multi-omics breast cancer example, we have used multi-omics features of the same gene symbol as group structure. MISKmeans can also be applied to single omics data type with pathway database as group structure (i.e., a pathway targets a collection of genes, and two pathways may contain overlapping genes), which is often encountered in real data application. We apply MISKmeans to integrate three leukemia transcriptomic datasets, including Balgobind et al. (2011); Kohlmann et al. (2008); Verhaak et al. (2009) (details see **Table S1(b)**). We used BioCarta pathway (http://www.broadinstitute.org/gsea/msigdb/collections.jsp#C2) as prior group structure (BioCarta contains 217 pathways). In this multi-cohort leukemia example, there are three underlying subtypes defined by translocation or inversion of chromosomes: inv(16) (inversions in chromosome 16), t(15; 17) (translocations between chromosome 15 and 17), t(8; 21) (translocations between chromosomes t(8; 21), which have been well studied with different treatment responses and prognosis outcomes. We chose the tuning parameter $K = 3$ to be consistent with the underlying truth. The expression data for Verhaak, Balgobind range from [3.169, 15.132] while Kohlmann ranged from [0,1]. All the datasets were downloaded directly from the NCBI GEO website (GSE6891, GSE17855, GSE13159 respectively). There were 54,613 probe sets in each study, and we removed probe sets with any missing value in it. If multiple microarray probes matched to the same gene symbol, we selected the representative probe with the largest interquartile range (IQR) (Gentleman et al., 2006). We ended up with 20,154 unique genes across all three cohorts.

3.3.1. Main result.

In order to determine number of clusters K from the data, we selected 500 features of largest variance to perform gap statistics using the sparse Kmeans as clustering algorithm, in each of the three studies. The resulting gap statistics (**Figure S5**) implies $K = 3$ is optimal, which is consensus among all three studies. The resulting transcriptomic profiles of these three leukemia data sets after applying MISKmeans are shown in **Figure S6**. A common set of intrinsic genes are selected across three studies. The clustering patterns are distinct and consistent across three studies, with ARI (comparing to the underlying truth) equal to 0.893, 1, and 0.948 respectively.

Table 3. Leukemia data feature selection (numGenes: number of genes selected) and clustering accuracy (ARI in each study). Time is evaluated in minute.

types	nGenes	Verhaak	Balgobind	Kohlmann	time
MISKmeans	1001	0.89	1	0.95	1.5
MSKM	1009	0.89	0.96	0.95	0.72
ISKmeans Verhaak	987	0.93	NA	NA	0.34
ISKmeans Balgobind	1005	NA	0.79	NA	0.32
ISKmeans Kohlmann	1009	NA	NA	0.95	0.4

To further benchmark the performance, we also applied MetaSparseKmeans ignoring the pathway knowledge and ISKmeans in three individual studies separately, using the same BioCarta pathway as prior group structure. On the one hand, the clustering accuracy shown in Table 3 indicates that MISKmeans generally has slightly better clustering accuracy than MetaSparseKmeans and ISKmeans. On the other hand, MISKmeans and MetaSparseKmeans provide a unified feature selection results across three studies while ISKmeans generates unstable feature selection across studies with low Jacaard Indexes (0.344 for Balgobind vs. Kohlmann, 0.365 for Verhaak vs. Balgobind and 0.365 for Verhaak vs. Kohlmann). In this example, the computing time is very fast for all methods.

To further evaluate functional annotation of the selected genes by each method, we employed pathway enrichment analysis via Fisher's exact test using, BioCarta, Kegg and Reactome as three different testing pathway databases. Five methods, including MISKmeans (BioCarta), MetaSparseKmeans, ISKmeans on Verhaak (BioCarta), ISKmeans on Balgobind (BioCarta), and ISKmeans on Kohlmann (BioCarta) are compared, where (BioCarta) indicates the method utilized BioCarta as group structure. The jittered plots of $-\log_{10} p$-values are shown in Figure 3. MISKmeans and ISKmeans show more significant pathways consistently in different testing pathway databases, which is expected since we used BioCarta pathway as prior knowledge to guide feature selection. This indicates incorporating prior knowledge indeed improves feature selection in the sense that the selected feature is more biologically meaningful. Remarkably, MISKmeans can identify more significant pathways than ISKmeans in general, though both of these methods utilize prior biological knowledge. This indicates MISKmeans is

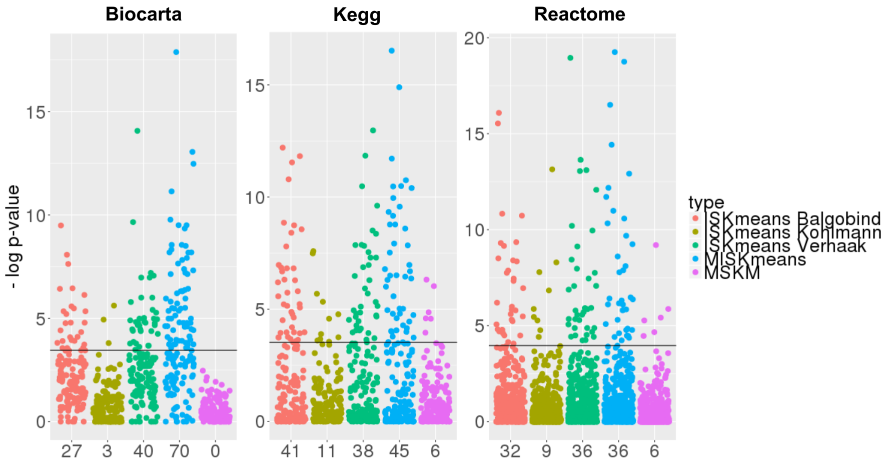

Figure 3. Pathway enrichment analysis of the five methods using BioCarta (left), Kegg (middle) and Reactome (right) as testing pathway database. Five methods, including MISKmeans (BioCarta), MetaSparseKmeans, ISKmeans on Verhaak (BioCarta), ISKmeans on Balgobind (BioCarta), and ISKmeans on Kohlmann (BioCarta) were compared, where (BioCarta) indicates the method utilized BioCarta as group structure. The horizontal black line is the 5% Bonferroni correction criteria. The number of pathways passing Bonferroni correction (correct p < 5%) is listed below the plot.

more powerful in selecting reliable and meaningful features than ISKmeans by combining multiple cohorts. Note that there is no overfitting issue when the testing pathway databases are Kegg and Reactome since they are different from the prior information Biocarta. Similarly, the results using Kegg and Reactome as prior information are in **Figure S7**.

3.3.2. Robustness analysis.

We performed robustness analysis to assess the performance of MISKmeans (Biocarta), ISKmeans (Biocarta) and MetaSparseKmeans. For robustness analysis, we randomly selected 2/3 samples in each study and applied the three methods respectively in each iteration. We repeated the above interations 100 times. To benchmark the feature selection robustness, we calculated all pairwise Jacaard index of these 100 iterations – totally $\binom{100}{2} = 4,950$ Jacaard index. Note that similar to the previous analysis, we chose the tuning parameter such that approximately 1,000 features are selected. To benchmark the clustering assignment robustness, we calculated the ARI between each iteration and the underlying clustering label. Note that we didn't perform pairwise ARI between each iteration because the selected samples may differ too much and thus the resulting ARI may not reliable. The mean and standard deviation of these Jacaard indexes are shown in **Table S6**. We observed that MISKmeans (Biocarta) and MetaSparseKmeans are more robust than ISKmeans (Biocarta) in terms of feature selection and clustering accuracy.

3.3.3. Cross Validation.

We performed 3-folds cross validation to evaluate the prediction performance of these methods. To be specific, we roughly split each study into 2/3 training dataset and 1/3 testing dataset. We performed MISKmeans (Biocarta), ISKmeans (Biocarta) and

MetaSparseKmeans in the training dataset respectively, such that approximately 1,000 features are selected for each method. Then we performed weighted Kmeans on the testing dataset using the selected features. The resulting clustering assignment was compared to the underlying truth using ARI by averaging the results from the three folds. The result (**Table S7**) showed again that MISKmeans (Biocarta) and MetaSparseKmeans are better than ISKmeans (Biocarta) in terms of prediction accuracy.

Discussion and Conclusion

Many diseases are heterogeneous, with many subtypes that differ by response to treatment, survival and biological pathways. Disease subtype discovery is an essential step in delivering personalized medicine. Subtype identification is usually labor intensive and requires combined expertise from oncologists and pathologists. Recently subtype discovery via omics data has become a popular approach, but still, reproducibility and cross-validation are big issues due to the heterogeneity of different cohorts and types of omics data. Hence in this manuscript, we propose two-way omics data integration via combining multi-cohort, multi-omics data and prior biological knowledge, which is practical and appealing since abundant omics datasets are available in public databases and repositories. Disease subtyping via two-way omics data integration conceptually generates the most comprehensive and reliable subtype definitions. The superior performance has been demonstrated in the multi-cohort and multi-omics breast cancer dataset and leukemia dataset. The proposed method also has demonstrated better performance compared with MetaSparseKmeans and ISKmeans.

MISKmeans has the following innovations. Firstly, this paper is among the first to propose the concept for simultaneous multi-cohort and multi-omics data integration and incorporating established biological knowledge in subtype discovery field. Meta-analytic framework of disease subtype analysis and multi-omics integrative clustering analysis are powerful approaches to identify disease subtypes; our proposed MISKmeans closes the literature gap between them and provides the most comprehensive characterization of disease subtypes. Secondly, the prior biological knowledge can be incorporated via the overlapping group lasso penalty. Fully accounting for the inter-omics regulatory relationship and external biological information increases interpretation of feature selections. Thirdly, the optimization of the objective function is very challenging since it involves iteratively updating weight, clustering assignment and pattern matching. It is solved by adopting the alternating direction method of multipliers (ADMM).

MISKmeans can also accommodate missing data. The first type of missingness is on the study level, in which one type of omics profile for a particular study is totally missing. In this scenario, we can revise Equation 4 to remove the contribution of the type of omics profile for the particular study and reweigh the contribution of other studies. The MISKmeans algorithm still leads to valid subtyping results based on the rest of omics profile. The second type of missingness is within one type of omics profile (e.g., some samples are missing for certain genes within one type of omics profile). This won't affect the effectiveness of MISKmeans since Equation 4 only depends on between cluster sum of square (BCSS) and total sum of square (TSS), which are still computable.

Our omics integration framework can be extended toward different distance metrics for other types of omics data. For example, the Bray-Curtis dissimilarity can be used for for count data, and Jaccard distance can be used for binary variables. A common penalty γ was adopted for all omics types in Equation 3 and Equation 4. Since each feature is standardized by TSS so its separation ability is directly comparable, which allows a fair competition between different omics types. However, if one of the two types of data is of lower quality or has a much smaller number of measured features, such penalty design may be in favor of the features of better quality and the omics type with larger number of measured features. In order to circumvent this issue, users can further extend the method and introduce an omics type specific penalty γ_t, where t is the omics type index. By tuning γ_t, users will have the flexibility to incorporate some subjective beliefs into the MISKmeans objective function such that the resulting subtypes won't be dominated by certain omics types. For example, the users can set equal γ_t if they want equal contribution of the different omics types. Or set γ_t proportional to the number of measured features of each omics type, in order to penalize more on the omics type with larger number of input features.

The computing for MISKmeans can be decomposed as three components based on the optimization procedure: weighted Kmeans, pattern matching and weight updating. Firstly, the convergence of weighted Kmeans is fast, which is a particular version of the classification EM algorithm. Secondly, the pattern matching complexity is reduced from $(K!)^{S-1}$ to $S(S-1)K!/2$ by caching intermediate results, which make moderate scale subtype analysis more feasible. Thirdly, we adopt ADMM to perform weight updating via adaptive augmented Lagrange parameters, which makes ADMM converge much faster. We further use C++ inside R software to accelerate ADMM. In fact, it will only take about 0.5 hour for MISKmeans when applying to the breast cancer example with in total 16,456 features, 7,989 groups, 5 subtypes and 1,765 samples across two cohorts. Note that the computing is done on a regular computer with a single AMD Opteron(tm) Processor (1.4GHz).

However, MISKmeans may suffer the limitation of scalability. For example, when K is very large (e.g. K = 10 and S = 2), the matching complexity will still goes to $S(S-1)K!/2 = 3,628,800$, which will be a big burden for both memory and computing time.

The current clustering methodology for two-way horizontal and vertical omics integration framework is based on Kmeans algorithm. Such framework can also be based on other clustering methods, including Gaussian mixture models, Bayesian nonparametric methods, etc. The Kmeans based methods usually have good performance when the underlying clusters do not overlap. However, when the underlying clusters overlap, Gaussian mixture model based methods may be expected to outperform Kmeans based methods. The idea of meta-analytic multi-omics data integration can be further extended to other aspects of statistical genomics, such as classification, association, dimensional reduction and network analysis, to generate a better understanding and interpretation of complex omics datasets. Since the current paper focused on the clustering algorithm in order to identify disease subtypes, we didn't consider the complex relationship between different layers of omics data in our simulation. For example, methylation levels in promoter regions are found negatively correlated with gene expression levels (Bell et al., 2011); and methylation levels in gene body are found positively correlated with gene expression levels (Jones, 1999, 2012). However, these inter-omics relationship should be consider in simulations when the goal is to identify associations or causal relationships between different layer of omics data.

Data Availability Statement

An efficient R package calling C++ is publicly available on Github (https://github.com/Caleb-Huo/MIS-Kmeans). All evaluation scripts for the simulation, the Leukemia data and the breast cancer example are on GitHub https://github.com/Caleb-Huo/MISKmeansSupp script folder. The Leukemia data are available under GEO (GSE6891, GSE17855, GSE13159), The TCGA breast cancer data is available `https://portal.gdc.cancer.gov/projects/TCGA-BRCA`, The METABRIC breast cancer data is available `https://www.synapse.org/#!Synapse:syn1688369`.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

H.Z., L.Z, T.M. and L.H participated in the algorithm development and results evaluation. H.S., L.D, Z.J, and G.T. provided the idea, structure and datasets for this manuscript. All authors participated in the writing of the manuscript.

Acknowledgements

L.Z., T.M. and G.T. are supported by NIH R01CA190766. D.L. is supported by grants from Bankhead-Coley Cancer Research Program and James and Esther King Biomedical Research Program, Florida Department of Health (6BC03 and 6JK03) and the Florida Breast Cancer Foundation.

References

Abramson, V. G., Lehmann, B. D., Ballinger, T. J., and Pietenpol, J. A. (2015). Subtyping of triple-negative breast cancer: Implications for therapy. *Cancer*, 121(1):8–16.

Balgobind, B. V., Van den Heuvel-Eibrink, M. M., De Menezes, R. X., Reinhardt, D., Hollink, I. H., Arentsen-Peters, S. T., van Wering, E. R., Kaspers, G. J., Cloos, J., de Bont, E. S., et al. (2011). Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica*, 96(2):221–230.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome biology*, 12(1):R10.

Bottolo, L., Chadeau-Hyam, M., Hastie, D. I., Zeller, T., Liquet, B., Newcombe, P., Yengo, L., Wild, P. S., Schillert, A., Ziegler, A., et al. (2013). Guess-ing polygenic associations with multiple phenotypes using a gpu-based evolutionary stochastic search algorithm. *PLoS genetics*, 9(8):e1003657.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Bredesen, D. E. (2015). Metabolic profiling distinguishes three subtypes of alzheimer's disease. *Aging*, 7(8):595–600.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl_1):i84–i90.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

Domany, E. (2014). Using high-throughput transcriptomic data for prognosis: A critical overview and perspectives. *Cancer Research*, 74(17):4612–4621.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210.

Fan, X. and Kurgan, L. (2015). Comprehensive overview and assessment of computational prediction of microrna targets in animals. *Briefings in bioinformatics*, 16(5):780–794.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Science & Business Media.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42.

Huo, Z., Tseng, G., et al. (2017). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The Annals of Applied Statistics*, 11(2):1011–1039.

Ivshina, A., George, J., Senko, O., Mow, B., Putti, T., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, 66(21):10292.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Jacob, L., Obozinski, G., and Vert, J. P. (2009). Group lasso with overlap and graph lasso. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440. ACM.

Jones, P. A. (1999). The dna methylation paradox. *Trends in genetics*, 15(1):34–37.

Jones, P. A. (2012). Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley. com.

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*.

Kim, S., Xing, E. P., et al. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *The Annals of Applied Statistics*, 6(3):1095–1117.

Kodama, Y., Shumway, M., and Leinonen, R. (2011). The sequence read archive: explosive growth of sequencing data. *Nucleic acids research*, 40(D1):D54–D56.

Kohlmann, A., Kipps, T. J., Rassenti, L. Z., Downing, J. R., Shurtleff, S. A., Mills, K. I., Gilkes, A. F., Hofmann, W.-K., Basso, G., DellOrto, M. C., et al. (2008). An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in leukemia study prephase. *British journal of haematology*, 142(5):802–807.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of clinical investigation*, 121(7):2750.

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, page bbw113.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J. A., Klijn, J. G., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M. J., and Sotiriou, C. (2007). Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology*, 25(10):1239–1246.

Lu, S., Li, J., Song, C., Shen, K., and Tseng, G. C. (2009). Biomarker detection in the integration of multiple multi-class genomic studies. *Bioinformatics*, 26(3):333–340.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8):1160–1167.

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C.-H., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.-M., Gallia, G. L., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807–1812.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752.

Quintana, M. and Conti, D. (2013). Integrative variable selection via bayesian model uncertainty. *Statistics in medicine*, 32(28):4938–4953.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine*, 5(9):e184.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annual review of statistics and its application*, 3:181–209.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–1947.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., et al. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature medicine*, 19(5):619–625.

Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., Ladanyi, M., and Sander, C. (2012). Integrative subtype discovery in glioblastoma using icluster. *PloS one*, 7(4):e35236.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal* *of Computational and Graphical Statistics*, 22(2):231–245.

Simon, R. (2005). Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute*, 97(12):866–867.

Simon, R., Radmacher, M. D., Dobbin, K., and McShane, L. M. (2003). Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1):14–18.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lonning, P. E., and Borresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874.

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C. M., Lonning, P. E., Brown, P. O., Borresen-Dale, A.-L., and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences*, 100(14):8418–8423.

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). Incorporating biological information into linear models: A bayesian approach to the selection of pathways and genes. *The annals of applied statistics*, 5(3).

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., Johnson, D. S., Trivett, M. K., Etemadmoghadam, D., Locandro, B., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208.

Tseng, G., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536.

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*. *Cancer cell*, 17(1):98–110.

Verhaak, R. G., Wouters, B. J., Erpelinck, C. A., Abbas, S., Beverloo, H. B., Lugthart, S., Löwenberg, B., Delwel, R., and Valk, P. J. (2009). Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *haematologica*, 94(1):131–134.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120.

Williams-Gray, C. H. and Barker, R. A. (2017). parkinson disease: Defining pd subtypesa step toward personalized management? *Nature Reviews Neurology*, 13(8).

Witkos, T., Koscianska, E., and Krzyzosiak, W. (2011). Practical aspects of microrna target prediction. *Current molecular medicine*, 11(2):93–109.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., Sotiriou, C., Oesterreich, S., and Tseng, G. C. (2016). Metadcn: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics*, 33(8):1121–1129.