Bioinformatics, 35(10), 2019, 1653–1659
doi: 10.1093/bioinformatics/bty866
Advance Access Publication Date: 17 October 2018
Original Paper

OXFORD

Genome analysis

Integrative cancer patient stratification via subspace merging

Hao Ding[1,†,‡], Michael Sharpnack[2,†], Chao Wang[1,2,§], Kun Huang[1,2,*,¶] and Raghu Machiraju[1,2,*]

[1]Department of Computer Science and Engineering and [2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
[‡]Present address: Uber Technologies Inc., San Francisco, CA 94103, USA
[§]Present address: Thermo Fisher Scientific, South San Francisco 94080, USA
[¶]Present address: Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

Received on November 18, 2017; revised on June 9, 2018; editorial decision on October 7, 2018; accepted on October 15, 2018

Abstract

Motivation: Technologies that generate high-throughput omics data are flourishing, creating enormous, publicly available repositories of multi-omics data. As many data repositories continue to grow, there is an urgent need for computational methods that can leverage these data to create comprehensive clusters of patients with a given disease.

Results: Our proposed approach creates a patient-to-patient similarity graph for each data type as an intermediate representation of each omics data type and merges the graphs through subspace analysis on a Grassmann manifold. We hypothesize that this approach generates more informative clusters by preserving the complementary information from each level of omics data. We applied our approach to The Cancer Genome Atlas (TCGA) breast cancer dataset and show that by integrating gene expression, microRNA and DNA methylation data, our proposed method can produce clinically useful subtypes of breast cancer. We then investigate the molecular characteristics underlying these subtypes. We discover a highly expressed cluster of genes on chromosome 19p13 that strongly correlates with survival in TCGA breast cancer patients and validate these results in three additional breast cancer datasets. We also compare our approach with previous integrative clustering approaches and obtain comparable or superior results.

Availability and implementation: https://github.com/michaelsharpnack/GrassmannCluster

Contact: kunhuang@iu.edu or machiraju.1@osu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the past decade, the large consortium efforts of The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium have pushed the boundaries of personalized medicine. The prior understanding of tumor subtypes based on histology and immunohistochemical markers has been complicated by the vast amounts of high-throughput data available. While many of the previous classifications based on clinical attributes are still useful, methods that leverage the omics data produced by TCGA have the potential to develop new clinically useful biomarkers and investigate tumor biology simultaneously. This data can be used individually to stratify tumors biologically and clinically; however, doing so is not without its challenges. Methods that group tumors based on this high dimensional data must separate useful signals from thousands of noisy measurements. The problem is further compounded when

integrating all of the diverse data types, such as microRNA, mRNA and DNA methylation, available for each tumor.

Methods used to leverage multiple sources of high dimensional biological fall into three broad categories: early integration, late integration and intermediate integration. Methods with early integration fuze all data types into a single dataset and perform analytical methods directly on it (Fridley *et al.*, 2012; Mankoo *et al.*, 2010). In late integration, separate models are applied independently to each data type and the integration is conducted by assembling the results at the end (Cancer Genome Atlas Research Network *et al.*, 2012; Verhaak *et al.*, 2010). Intermediate integration combines multiple data types after transforming each data type into an intermediate representation (e.g. a graph or a kernel matrix) (Huang *et al.*, 2012; Shen *et al.*, 2009; Speicher and Pfeifer, 2015; Wang *et al.*, 2014).

Our proposed approach falls into the intermediate integration paradigm. Given multiple types of omics data for the same set of patients, we create a patient-to-patient similarity graph for each data type as an intermediate representation and merge the graphs through subspace analysis on a Grassmann manifold. This avoids problems created by using data types with differing numbers of measurements, as is the case in genomic data. The resulting combination can then be viewed as a lower-dimensional representation of the original data. Subspace analysis on a Grassmann manifold has been previously studied in the context of computer vision (Dong *et al.*, 2014) but has not been used for integrating genomic data. Finally, we cluster the patients on this lower-dimensional subspace to identify potential tumor subtypes.

Our approach has several advantages over the early and late integration: (i) intermediate representation preserves data-type-specific properties; (ii) the intermediate integration approach is robust to different data measurement scales; (iii) this approach can be used to integrate many types of data, including continuous or categorical values, as long as the data contain a unifying feature; (iv) unlike previous approaches, such as similarity network fusion (SNF; Wang *et al.*, 2014) and affinity aggregation for spectral clustering (Huang *et al.*, 2012), our method does not involve iterative optimization. We timed our method on the TCGA breast cancer dataset using a personal computer (8 gb memory, 2.3 GHz processor) on Matlab, and found that it completed in 14 s. For results comparison, we applied our method to the datasets analyzed by a recent work (Wang *et al.*, 2014). Our survival analysis results are comparable or even better than those reported in Wang *et al.* (Wang *et al.*, 2014).

Next, we applied our method to the TCGA breast cancer tumor samples with matched RNA, microRNA and DNA methylation and clinical follow-up data. We chose breast cancer because there is a rich history of dividing breast cancer into clinically useful subtypes. RNA expression panels such as PAM50 (Perou *et al.*, 2000) and OncotypeDX (Jackisch *et al.*, 2009) use select genes to separate patients into prognostic, biological and therapeutically distinct groups. The success of the PAM50 to subtype breast cancer, in part, likely reflects the fact that luminal type breast cancers and triple negative breast cancers are molecularly distinct cancers that happen to develop in the same anatomical location. The utility of future classifiers of breast cancer patients hinges on their ability to further separate the four canonical subtypes of breast cancer into therapeutically and prognostically useful groups. For example, OncotypeDX uses measurements of 21 genes to calculate a recurrence score, which helps physicians and patients decide whether or not to use adjuvant chemotherapy. The interpretability of this test is limited by the existence of an intermediate result. We show that by integrating gene expression, microRNA, and DNA methylation data, we produce clinically useful subtypes of breast cancer, as well as investigate the molecular mechanisms underlying these subtypes. Specifically, we find that a group of patients with excellent prognosis shows high expression of a large cluster of genes located on chromosome 19p13. We were able to validate these findings on three additional datasets.

2 Materials and methods

2.1 Dataset and data pre-processing

We downloaded TCGA level 3 datasets containing gene expression, miRNA expression and DNA methylation expression profiles from 441 primary tumors of breast cancer patients. Multiple platforms for each data type are available via TCGA. We chose UNC-Illumina-Hiseq-RNASeq platform for gene expression, BCGSC-Illumina-Hiseq-miRNAseq platform for miRNA expression and JHU-USC-Human-Methylation-450 k platform for DNA methylation expression profiles. Pertinent clinical data were also available for all of these patients. The minimum follow-up duration is 3 months (91 days), and the median follow-up duration is 16 months (492 days). We performed the following normalization for each data type.

$$\hat{f} = \frac{f - E(f)}{\sqrt{Var(f)}} \tag{1}$$

where f is a feature any data type, $\hat{f}$ is the corresponding feature after normalization and $E(f)$ and $Var(f)$ represent the sample mean and sample variance of f, respectively.

To validate our findings in breast cancer, we downloaded two datasets from the NCBI Gene Expression Omnibus with the accession numbers GSE3143 and GSE1456. The NKI breast cancer gene expression data was downloaded from ccb.nki.nl.

2.2 Construction of the patient-to-patient similarity graph

Consider M types of omics data measurements $\{\mathbf{X}^m\}_{m=1}^M$ (each with dimension of p_m) collected from N patient samples, such that $\mathbf{X}^m$ is a $p_m \times N$ matrix. For each data type $\mathbf{X}^m$, we construct a patient-to-patient similarity graph G^m to model the local neighborhood relationships between the samples. Let $G^m = (V^m, E^m, W^m)$ denote a patient similarity graph for data type m, where V^m represents the vertex set, E^m represents the edge set and W^m represents the adjacency matrix.

The adjacency matrix W^m of the graph G^m is a symmetric matrix whose entry w_{ij}^m represents the edge weight if there is an edge between vertex v_i and v_j, or 0 otherwise. To construct this similarity graph, we first compute a similarity matrix to measure the pairwise similarity between each sample pair. Here, we use a heat kernel as the similarity metric:

$$S_{ij}^m = exp\left(-\frac{\|x_i^m - x_j^m\|^2}{2t^2}\right), i = 1, ..., N, j = 1, ...N. \tag{2}$$

We then extract k-nearest neighbors graph from the similarity matrix S^m: We denote $\mathbf{N}_i$ as a set of v_i's neighbors including v_i and size of $\mathbf{N}_i$ is k. The number of k normally depends on the sample size. We then connect v_i and v_j with an undirected edge with edge weight as S_{ij} if $v_j \in N_i$ [Equation (3)].

$$W_{ij}^m = \begin{cases} S_{ij}^m, & \text{if } v_j \in \mathbf{N}_i. \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Essentially we make the assumption that local similarities are more reliable than remote ones. This is a mild assumption widely adopted by other manifold learning algorithms.

2.3 Construction of subspace representation

For each similarity graph G^m constructed from data type m, we first compute the normalized graph Laplacian matrix L^m, which is defined as $L^m = D^{m-\frac{1}{2}}(D^m - W^m)D^{m-\frac{1}{2}}$, where W^m and D^m denote the adjacency matrix and degree matrix of G^m, respectively.

Once obtain the normalized graph Laplacian matrix from each data type, we conduct the graph embedding to construct the subspace representation for each graph. The main purpose of graph embedding is to find a low-dimensional subspace that preserves similarities between the vertex pairs. In other words, the resulting subspace best captures the characteristics of each data types. We use U^m to denote the subspace representation of G^m. The optimal graph embedding in k dimension is derived by minimizing the following objective function:

$$\min_{U^m \in R^{N\times k}} tr(U^{m\prime} L^m U^m), \quad \text{s.t. } U^{m\prime} U^m = I. \tag{4}$$

By the Rayleigh–Ritz theorem (Horn and Johnson, 1985), the solution to the problem of Equation (4) is given by the first k eigenvectors of the Laplacian L^m, which can be computed using efficient algorithms for eigenvalue problems.

2.4 Merge subspace representations on Grassmann manifold

With the subspace representations $U^M_{m=1}$ construct from each data type, we then merge them on a Grassmann manifold. The Grassmann manifold is defined as a set of linear subspaces of a Euclidean space, therefore each subspace representation. We use $G(k, n)$ to denote a Grassmann manifold with k-dimensional linear subspaces in n dimensional Euclidean space R^N. An element of $G(k, n)$ can be represented by an orthonormal matrix $Y \in R^{n\times k}$ whose columns span the corresponding k-dimensional subspace in R^n; it is thus denoted as $span(Y)$. A distance between two subspaces is defined as the length of the shortest geodesic connecting the two corresponding points on the Grassmann manifold. However, there is a more convenient and efficient way of defining distances using the projection distance (Golub and Van Loan, 2012). For instance, the projection distance between $span(Y_1)$ and $span(Y_2)$ $(Y_1, Y_2 \in n \times k)$ is defined as follow:

$$\begin{aligned} d^2_{proj}(Y_1, Y_2) &= \sum_{i=1}^{k} \sin^2\theta_i \\ &= k - \sum_{i=1}^{k} \cos^2\theta_i \\ &= k - tr(Y_1 Y_1{}' Y_2 Y_2{}'). \end{aligned} \tag{5}$$

With the distance measurement defined above, we can capture the similarity between the subspaces on Grassmann manifold and subsequently enable us to merge the information from multiple graphs in a meaningful manner. In the last section, we constructed subspace representations $\{U^m\}^M_{m=1}$ for M data type by the conduct the spectral embedding on the patient-to-patient similarity graphs $\{G^m\}^M_{m=1}$. Each subspace representation $\{U^m\}$ defines a k-dimensional subspace in R^n, where n is the number of patients and k is the target number of clusters. To merge these subspaces, we want to find an integrative subspace $span(U)$, which is close to all the individual subspaces $span(U^m)$, and at the same time the representation U preserves the vertex connectivity in each G^m.

With the distance measurement defined in Equation (5), we can define a summation of projection distance between the integrative subspace U and M subspaces $\{U^m\}^M_{m=1}$ as follow:

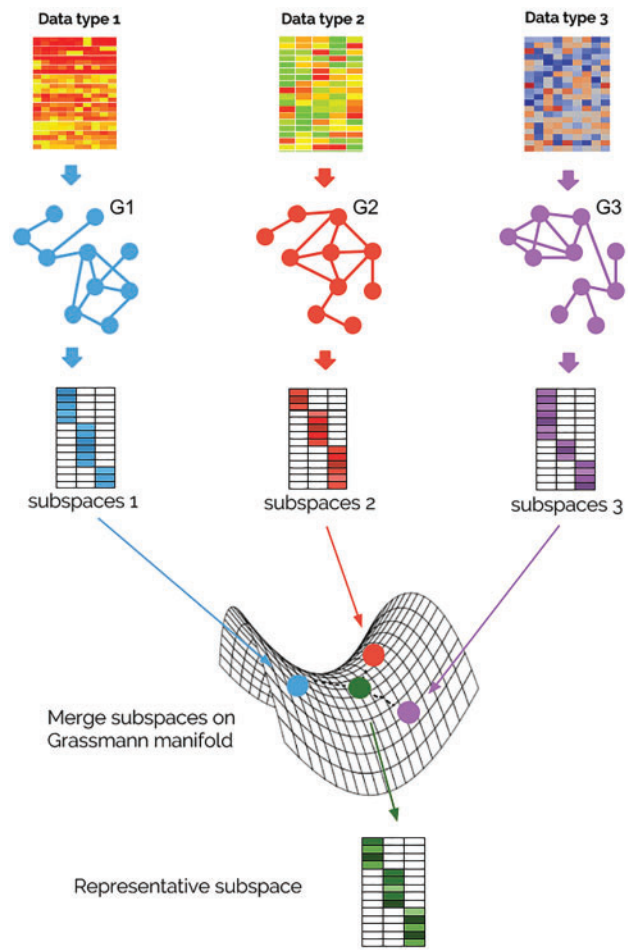

Fig. 1. Workflow for integrating and merging cancer genomics datasets. A patient-to-patient similarity graph is constructed for each data type. The similarity matrices are converted to subspaces and embedded in a Grassmann manifold, where they are integrated into a single, representative subspace. This subspace is then clustered to obtain the final integrative patient groups

Table 1. Comparison of Cox survival *P*-values from integrative clustering on a Grassmann manifold with those from SNF

Cancer type	SNF (nature.2014)	Our method
GBM (3 clusters)	2.0×10^{-4}	4.3×10^{-3}
BIC (5 clusters)	1.1×10^{-3}	2.0×10^{-4}
KRCCC (3 clusters)	2.9×10^{-2}	2.8×10^{-2}
LSCC (4 clusters)	2.0×10^{-2}	1.6×10^{-2}

$$\begin{aligned} d^2_{proj}(U, \{U^m\}^M_{m=1}) &= \sum_{m=1}^{M} d^2_{proj}(U, U^m) \\ &= \sum_{m=1}^{M} [k - tr(UU'U^m U^{m\prime})] \\ &= kM - \sum_{i=1}^{M} tr(UU'U^m U^{m\prime}). \end{aligned} \tag{6}$$

The subspace U which minimizes Equation (6) is close to all the individual subspaces $\{U^m\}^M_{i=1}$ in terms of the projection distance on the Grassmann manifold. Since we also want U to preserve the vertex connectivity in graphs from each data type. Therefore, we finally propose to merge multiple subspaces by solving the following optimization problem that integrates Equations (4) and (6):

$$\min_{U \in R^{n \times k}} \sum_{m=1}^{M} tr(U'L^m U) + \alpha[kM - \sum_{m=1}^{M} tr(UU'U^m U^{m\prime})], \quad \text{s.t. } U'U = I \tag{7}$$

where L^m and U^m are the graph Laplacian and the subspace representation for G^m, respectively. Such a representation not only preserves the structural information contained in the individual graph, which is encouraged by the first term of the objective function in Equation (7), but also keeps a minimum distance between itself and the multiple subspaces, which is enforced by the second term. The regularization parameter α balances the trade-off between the two terms in the objective function. By ignoring constant terms and rearranging the trace form in the second term of the objective function, Equation (7) can be rewritten as:

$$\min_{U \in R^{n \times k}} tr[U'(\sum_{i=1}^{M} L^m - \alpha \sum_{m=1}^{M} U^m U^{m\prime})U], \quad \text{s.t. } U'U = I. \tag{8}$$

The Equation (8) shares the same form with Equation (4), and the solution to the problem of Equation (8) is then the first k eigenvectors of the modified Laplacian L_{mod} in Equation (9), followed by the Rayleigh–Ritz theorem (Horn and Johnson, 1985).

$$L_{mod} = \sum_{m=1}^{M} L^m - \alpha \sum_{m=1}^{M} U^m U^{m\prime}. \tag{9}$$

Finally, we can cluster resulting integrative subspace U using the k-means algorithm.

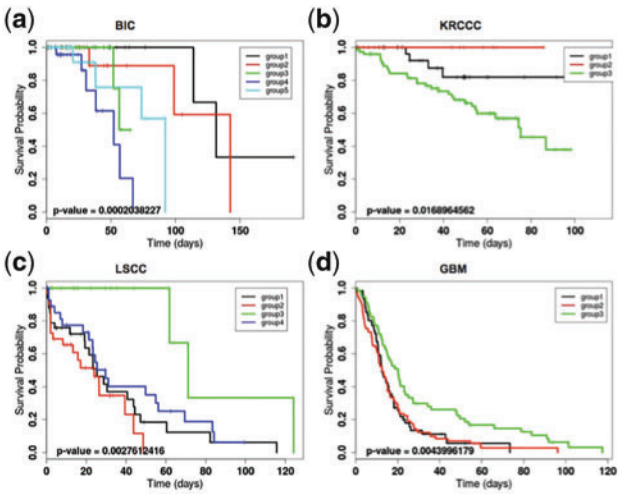

Fig. 2. Integrative clustering discovers clinically significant subtypes of cancer. Shown are Kaplan–Meier plots of the overall survival of integrative clusters for breast invasive carcinoma (BIC) (**a**), kidney renal clear cell carcinoma (KRCCC) (**b**), lung squamous cell carcinoma (LSCC) (**c**) and glioblastoma multiforme (GBM) (**d**). *P*-values are computed from the logrank test

2.5 Molecular basis of integrative clusters

We construct an integrated co-expression network by calculating the feature-to-feature spearman correlation. After filtering nodes for low expression and variance (removing the bottom quartile from each omics data type), and edges for low correlation (Spearman $\rho < 0.7$), we are left with a network composed of 9195 nodes and 257 703 edges.

Differential expression is calculated by a *t*-test with false discovery rate (FDR) correction. Given the large numbers of patients in each group and the strength of molecular differences between each group, this method, in lieu of more complicated methods, was deemed sufficient to detect a large portion of differential expression.

Genes contained in chr19p13 were downloaded from the molecular signatures database (http://software.broadinstitute.org/gsea/msigdb). Copy number variants (CNV) data for genes in chr19p13 were downloaded from cBioPortal (http://www.cbioportal.org) (Gao *et al.*, 2013).

2.6 Method overview

Given multiple types of omics data for the same set of patients, we first create a patient-to-patient similarity graph for each data type (Fig. 1). In our breast cancer case study, this means that three unique patient-to-patient graphs are created, one each for microRNA, methylation and mRNA expression data. Prior to any further analysis, we remove edges with low similarity measures, which represent uncertain relationships between patients. To perform the integrative analysis, the differences between these graphs must be reconciled. Prior studies employ an iterative convergence approach (Wang *et al.*, 2014); in our work, we merge the graphs in two steps via a mathematical construct, the Grassmann manifold.

In the first step, the structure of each patient-to-patient similarity graph is captured by a subspace representation via spectral embedding. Since each graph is reduced to its low-dimensional subspace representation, computation is minimal and noise is reduced. In the second step, each subspace representation is considered as a point on a Grassmann manifold. Therefore, we can find a new representative subspace on the Grassmann manifold where the overall distance between new representative subspace and the individual subspaces is minimized. The result of this analysis is a subspace that summarizes the desired merged graph, containing information from all data types.

Finally, we clustered the patients in the resulting representative subspace and conducted a *post hoc* analysis to evaluate the clustering results. We hypothesize that such a method will generate more informative clusters by preserving the complementary information from each level of omics data. For further details, see Materials and methods.

2.7 Comparison with results from similarity network fusion

To demonstrate the effectiveness of our methods, we applied our method on multiple datasets analyzed by Wang *et al.* (2014) and

Table 2. Clinical attributes of TCGA breast cancer subtypes

	Group 1 ($N = 83$)	Group2 ($N = 76$)	Group3 ($N = 103$)	Group4 ($N = 111$)	Group5 ($N = 68$)
ER, (+)	23 (27.7%)	55 (72.4%)	100 (97.1%)	103 (92.8%)	54 (79.4%)
(−)	59 (71.1%)	19 (25%)	3 (2.9%)	6 (5.4%)	13 (19.1%)
PR, (+)	19 (22.9%)	51 (67.1%)	88 (85.4%)	95 (85%)	46 (67.6%)
(−)	63 (75.9%)	23 (30.3%)	15 (14.6%)	13 (11.7%)	20 (29.4%)
HER2 (+)	4	9	11	12	7
(−)	46	41	50	62	40

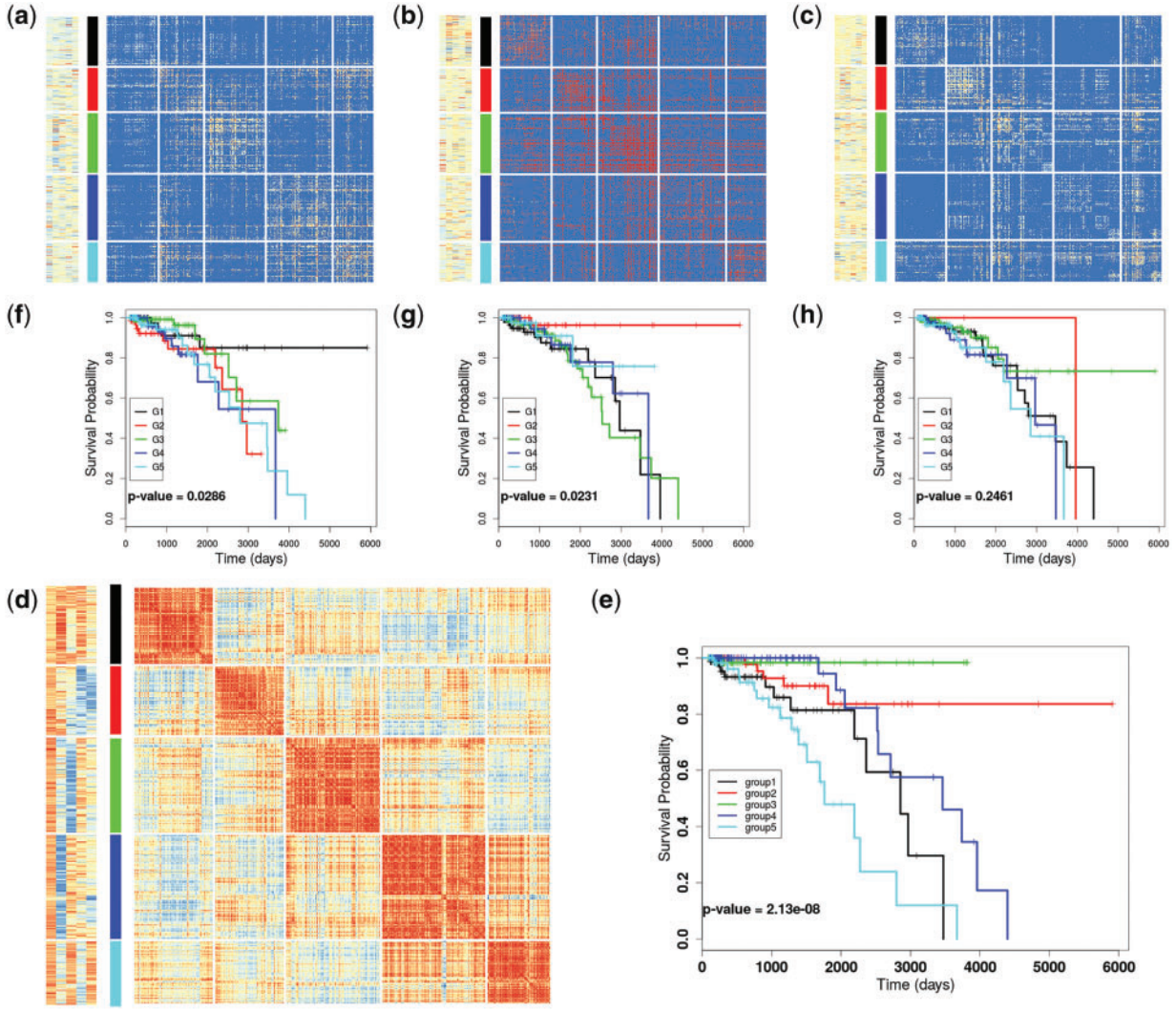

Fig. 3. Integrative clustering of breast tumors produces prognostically relevant and biologically significant groupings. (**a–c**) The adjacency matrices of patient-to-patient similarity graphs, produced from mRNA (a), microRNA (b) and methylation (c) datasets. (**d**) Integrative clustering of patients using all three datasets. Color bars at left show the clusters of patients, and the heatmap to the left of each colorbar shows the eigenvector clustering results. (**e–h**) Survival analysis of patient stratification results using integrative and single-data-type clustering methods. Kaplan–Meier survival curves of clusters produced by: (e) integrative clustering, (f) gene expression alone, (g) miRNA expression alone, (e) DNA methylation alone, listed along with estimated *P*-values (logrank test)

compared our results to those generated by their SNF method. The cancer types include glioblastoma multiforme (GBM, $N=213$), breast invasive carcinoma (BIC, $N=$), kidney renal clear cell carcinoma (KRCCC, $N=122$) and lung squamous cell carcinoma (LSCC, $N=106$). The Cox survival *P*-values reported by Wang *et al.* (2014) and our method are listed in Table 1. In order to ensure the comparability of the results, we inherit Wang *et al.*'s (2014) choice of cluster numbers for each cancer type. As can be seen in the table, in three out of the four types of cancer studied our method provides more significant differences between the survival times. In GBM, the *P*-value is comparable to the *P*-value generated by SNF. Survival plots for GBM, BIC, KRCCC and LSCC tumors are shown in Figure 2. Our proposed method can also be extended to improve the survival rate prediction task. We trained random survival forest with subspace representation generated from integrative subspace and individual data types integrative subspace, respectively (Supplementary Fig. S4) and found that our integrative approach displayed superior performance in comparison to using only a single data type.

2.8 Identification of integrative subtypes of breast cancer

For decades, researchers have been developing methods to subtype breast cancer. Gene signatures, such as PAM50 (Perou *et al.*, 2000) and the 70-gene signature (Van't Veer *et al.*, 2002) have been developed to subtype breast cancer patients. However, subtyping based on other data sources, e.g. microRNA and DNA methylation lead to incoherent results when compared to subtyping done using mRNA data (Blenkiron *et al.*, 2007; Stefansson *et al.*, 2015; Yuan *et al.*, 2011).

In this paper, we obtained DNA methylation, mRNA expression and miRNA expression data from 441 primary tumors of breast cancer patients from TCGA (Cancer Genome Atlas Research Network *et al.*, 2012). Detailed clinical information for this cohort is listed in Table 2. We applied the proposed method to this dataset and partitioned the patient population into five integrative subtypes. The choice of the number of clusters is determined by two factors: (i) the silhouette scores reached a peak when the number of cluster is 5 (Supplementary Figs S1 and S2); (ii) the survival analysis Cox

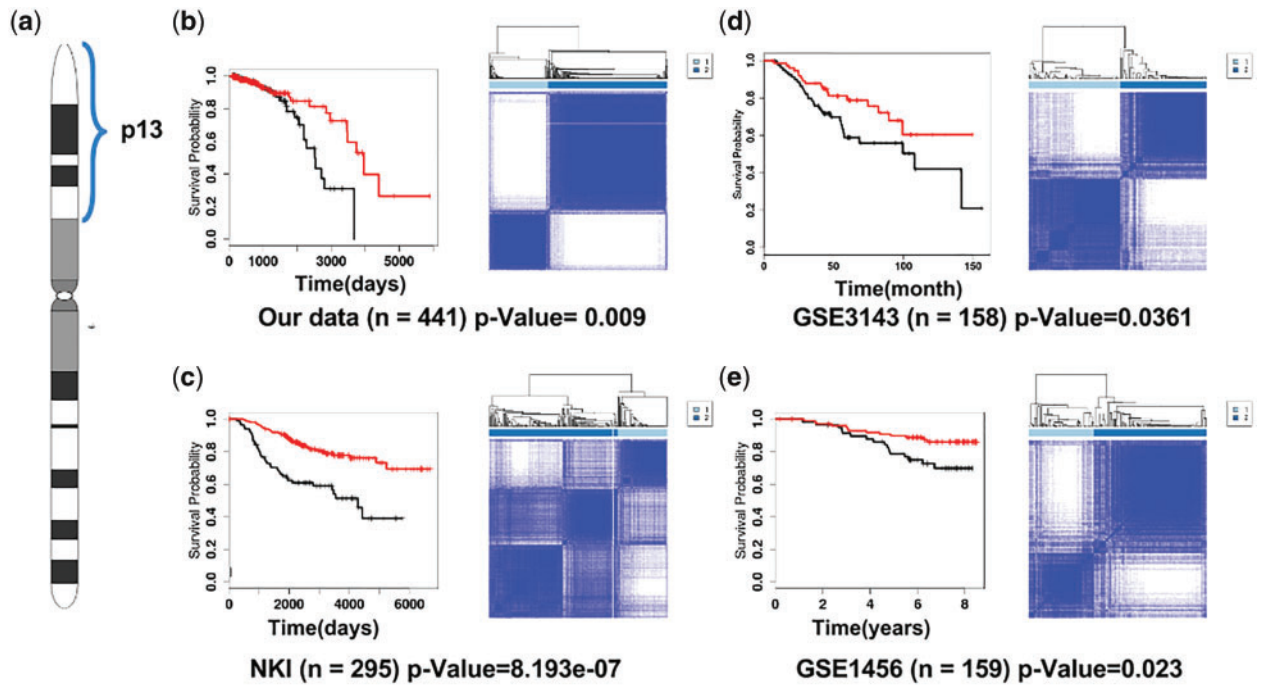

Fig. 4. Density of differentially expressed genes on chromosome 19p13 are shown in (**a**). (**b–e**) Survival curves and clustered heatmaps produced by separating four datasets based on expression of genes on chr19p13. Kaplan–Meier survival curves, clustered heatmaps and accompanying *P*-values (logrank test) generated from the training dataset from TCGA are shown in (**b**). Results generated from three validation sets are shown in (**c–e**): Netherlands Cancer Institute (NKI) in (**c**), GSE3143 in (**d**) and GSE1456 in (**e**)

P-values obtained for each cluster size found a global minimum at $k = 5$ (Supplementary Fig. S3).

Figure 3a–c shows the adjacency matrices and their corresponding representative subspace for the patient-to-patient similarity graphs from each data type. As shown in the plot, the connectivity of the graphs varies considerably from each other. For instance, group 2 (red) has high inner connectivity with the graph generated by miRNA and DNA methylation, while the graph constructed with mRNA data better supports the connectivity in group 4 (blue). Figure 3e–f compares clusters obtained from integrative versus single-data-type analyses. Figure 3a shows the overall survival plot produced by integrative clustering, with an estimated P-value < 0.0001 (logrank test). While analyses of mRNA (Fig. 3b) and miRNA (Fig. 3c) data independently yielded significant clustering results (logrank test $P = 0.0231$ and 0.0286, respectively), overall survival of the clusters is much more clearly separated using integrative clustering. Supporting this finding, the subspace representations shown in Figure 3a–d are most highly separated in the integrative cluster, shown in Figure 3d. The clustering produced from methylation data (Fig. 3d) was not significantly prognostic ($P = 0.2461$).

Subtype 1 (black) largely contains patients negative for progesterone (PR), estrogen (ER) and epidermal growth factor 2 (HER2) receptors, also known as triple negative (see Table 2). The vast majority of tumors in subtype 3 (green) and 4 (blue) are positive for ER and PR, but patients in subtype 3 have a clear survival advantage over patients in subtype 4. Similarly for groups 2 and 5, they have similar ER, PR and HER2 statuses, yet different prognoses. These differences may represent useful prognostic and, more importantly, therapeutic opportunities. To investigate the molecular basis of our subtyping, we further analyzed the differences in mRNA, microRNA and methylation abundances.

2.9 Molecular basis of integrative breast cancer subtypes

We have shown that our method produces subtypes with clinically relevant prognoses, yet of equal importance are the molecular alterations associated with these differences in prognoses. To assess the biological significance of our clustering results, we look for differentially activated groups of nodes in an integrative co-expression network, where each node is an mRNA gene, methylation site or microRNA species (see Supplementary Methods). Our first observation was that group 1, clinically identified as the triple-negative breast cancer (TNBC) subtype, indeed contains the molecular hallmarks previously associated with TNBC. For example, we observed that MYBL2, CENPA, AURKB and KIF2C are all overexpressed in group 1 (P-value < 0.001, FDR-corrected), which is consistent with previous results on the TNBC subtype (Sparano *et al.*, 2009) (Supplementary Fig. S5). Also consistent with prior studies, patients in group 1 have relatively poor overall survival.

Of particular interest are the molecular alterations that might cause tumors with similar hormone receptor statuses to have differing prognoses. For example, groups 3 and 4 have nearly identical hormone receptor statuses (Table 1), yet in contrast to group 4, no patients in group 3 die during the follow-up period. In our co-expression network we noticed that there was a large, highly coexpressed module of genes located on chr19p13 (Fig. 4a). Remarkably, a large subset of these genes were overexpressed in group 3 compared to other groups. We found that chr19p13 gene expression can clearly separate patients in the TCGA dataset into good and poor prognoses (Fig. 4b). To validate our findings, we further tested our 19p13 signature on three additional publicly available validation datasets [NKI (Van't Veer *et al.*, 2002), Fig. 4c, GSE3141 (Bild *et al.*, 2006), Fig. 4d and GSE1456 (Pawitan *et al.*, 2005), Fig. 4e]. Of note, all four datasets were created using different expression platforms, which shows

that our signature is platform independent. In addition, it has been reported that low SAFB, a gene present on chr19p13, expression is associated with worse outcome in breast cancer patients (Hammerich-Hille *et al*., 2010), which is consistent with our finding.

We further sought to investigate possible molecular alterations associated with chr19p13 overexpression. We first hypothesized that copy number changes in chr19p13 could explain the expression differences; however, an analysis of the corresponding CNV data showed that only a minority of the variation is explained by deletions or amplifications on chr19p13 (Supplementary Fig. S6).

Next, we searched the methylation data for an explanation, but there was no clear upregulation or downregulation of methyl sites located on chr19p13. Further experiments are necessary to identify a single gene, microRNA or methylation site responsible for the observed changes in chr19p13 expression. This search is complicated by the fact that numerous known tumor suppressors and oncogenes are located in this region. Another possible explanation is a large chromosomal event that is not easily detected in high-throughput sequencing data.

3 Conclusion

In this paper, we propose a novel method to perform efficient integrative patient stratification. Our approach aggregates information from multiple molecular expression data through a subspace analysis on the Grassmann manifold. We applied our method to stratify the breast cancer patient cohort datasets collected from TCGA by integrating gene expression, DNA methylation and miRNA expression data. The result demonstrates our method can leverage information from different omics data into clinically relevant subtypes. Also, through our subtyping results, we uncovered a group of genes located on chromosome 19p13 with strong prognostic power. We further validate our finding on three independent datasets. Future follow-up studies on this gene set are necessary to reveal its biological implication.

Since the integration step in our method is independent of the properties of data source, the input data types are not limited to genomic data. With appropriate similarity measurements, our method can be extended to applications in which integration of clinical categorical information and image datasets for which clustering is needed. Nonetheless, our approach can also be applied to other tasks that require integration of multiple types of features.

Funding

This work was supported by the following grants: National Cancer Institute [ITCR U01 CA188547-03]; Leidos [15X040]; Indiana University Precision Health Initiative; and The National Library of Medicine [T15LM011270].

Conflict of Interest: none declared.

References

Bild,A.H. *et al*. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Blenkiron,C. *et al*. (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.*, **8**, 10.

Cancer Genome Atlas Research Network et al. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519–525.

Dong,X. *et al*. (2014) Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Trans. Signal Process.*, **62**, 905–918.

Fridley,B.L. *et al*. (2012) A Bayesian integrative genomic model for pathway analysis of complex traits. *Gen. Epidem.*, **36**, 352–359.

Gao,J. *et al*. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, 269.

Golub,G.H. *and* Van Loan,C.F. (2012) *Matrix Computations*. JHU Press, Baltimore, MD.

Hammerich-Hille,S. *et al*. (2010) Low SAFB levels are associated with worse outcome in breast cancer patients. *Breast Cancer Res. Treat.*, **121**, 503–509.

Horn,R.A. *and* Johnson,C.A. (1985) *Matrix Analysis*. Cambridge University Press, Cambridge.

Huang,H. *et al*. (2012) Affinity aggregation for spectral clustering. *CVPR*, **2012**, 773–780.

Jackisch,C. *et al*. (2009) Evolution of the 21-gene assay oncotype DX[textregistered] from an experimental assay to an instrument assisting in risk prediction and optimisation of treatment decision-making in early breast cancer. *Eur. Oncol.*, **6**, 36–42.

Mankoo,P.K. *et al*. (2010) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*, **6**, e24709–e24709.

Pawitan,Y. *et al*. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.*, **7**, R953–R964.

Perou,C.M. *et al*. (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Shen,R. *et al*. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Sparano,J.A. *et al*. (2009) Genotypic characterization of phenotypically defined triple-negative breast cancer. *J. Clin. Oncol.*, **27**, 500.

Speicher,N.K. and Pfeifer,N. (2015) Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, **31**, i268–i275.

Stefansson,O.A. *et al*. (2015) A DNA methylation-based definition of biologically distinct breast cancer subtypes. *Mol. Oncol.*, **9**, 555–568.

Van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Verhaak,R.G.W. *et al*. (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–11.

Wang,B. *et al*. (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Yuan,Y. *et al*. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.