Quantitative Biology 2016, 4(1): 58–67
DOI 10.1007/s40484-016-0063-4

Review

Integrative clustering methods of multi-omics data for molecule-based cancer classifications

Dongfang Wang and Jin Gu*

Ministry of Education Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology/Department of Automation, Tsinghua University, Beijing 100084, China
* Correspondence: jgu@tsinghua.edu.cn

Received November 17, 2015; Revised January 13, 2016; Accepted January 23, 2016

One goal of precise oncology is to re-classify cancer based on molecular features rather than its tissue origin. Integrative clustering of large-scale multi-omics data is an important way for molecule-based cancer classification. The data heterogeneity and the complexity of inter-omics variations are two major challenges for the integrative clustering analysis. According to the different strategies to deal with these difficulties, we summarized the clustering methods as three major categories: direct integrative clustering, clustering of clusters and regulatory integrative clustering. A few practical considerations on data pre-processing, post-clustering analysis and pathway-based analysis are also discussed.

Keywords: clustering; cancer classification; omics; integrative analysis

INTRODUCTION

In current clinical practices, cancer is typically classified based on its tissue/cell-type origin and pathogens. As we know, cancer commonly involves complex molecular alterations. Molecule-based cancer re-classification and subtyping is becoming crucial for precision oncology [1–3]. Basically, molecule-based classifications come in two categories: supervised classification and unsupervised clustering. Supervised methods could discover genotype-phenotype interactions or patient risk stratifications based on labeled datasets (see a recent review in [4]). Unsupervised clustering plays a different role in terms of identifying novel cancer subtypes based on unlabeled molecular data. Unsupervised cancer subtyping has wide clinical applications and the molecule-based subtypes may be different with current clinical subtypes or stages mainly defined by pathologic features [5,6]. We will focus on unsupervised clustering methods in this review. Previous studies suggest that different molecular layers have different information for subtyping. For example, three surface receptors ER/HER2/PR principally define breast cancer subtypes which have been already used in clinical practices [7,8]; high mutation rate defines microsatellite instability (MSI) subtype [9]; CpG island hypermethylation defines CpG island methylator phenotype (CIMP) subtype [10]. As the rapid development of high-throughput biotechniques, such as next-generation sequencing techniques and high-density microarrays, we can simultaneously profile genomic, epigenomic and transcriptomic features of large-scale clinical samples with relatively low cost in short time. Several public cancer genome projects, such as International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA) and Asian Cancer Research Group (ACRG), have already released huge-volume cancer multi-omics data. One major challenge is how to integrate these data for better molecule-based cancer classification [11].

METHOD SURVEY

Overview: unsupervised clustering of cancer multi-omics data

Based on large-scale multi-omics data, unsupervised clustering is an important way to establish the landscapes

This article is dedicated to the Special Collection of Recent Advances in Next-Generation Bioinformatics (Ed. Xuegong Zhang).

of molecule-based cancer classifications. These methods should deal with two major difficulties. The first one is "data heterogeneity": different omics data have different data distributions and variation patterns, for example, sequencing data are usually modeled by count-based distributions, microarray data commonly use Gaussian distribution and mutation data follow binomial distribution. The other difficulty is "the complexity of inter-omics variations": different omics data have inter-layer regulatory co-variations (for example, gene expressions are regulated by copy number variations and promoter DNA methylations), but each layer also has its own specific variation patterns. Also, the methods should consider the curse of dimensionality because the number of molecular features is usually much larger than the sample size. According to the different ways to deal with above difficulties, we summarized the multi-omics integrative clustering methods as three major categories: i) direct integrative clustering; ii) clustering of clusters and iii) regulatory integrative clustering (Table 1).

Direct integrative clustering

One intuitive idea of integrating multi-omics data is to put all of them into a stacked matrix, and take this stacked matrix as input to the subsequent clustering analysis (Figure 1). To deal with the problem of data heterogeneity, this kind of method needs to process all types of data using a unified framework.

A simple strategy is to transform all types of data as relative scores by comparing the data from tumors and adjacent normal tissues [12]. Firstly, they used Cox regression to obtain prognosis-related features, including copy number variations, DNA methylations and gene expressions. Then, all the selected features were normalized and combined to a stacked matrix by comparing their values in each cancerous tissue against adjacent normal tissues. The final clusters were identified based on the stacked matrix using super *k*-means, which used Bayesian information criterion (BIC) score to automatically select the number of clusters.

Another strategy is to map different kinds of data to a shared low-dimension latent space. This strategy assumes there is a set of latent factors associated with a few driving oncogenic processes which generates the observed high-dimensional omics data. As an example, iCluster+ adopted this strategy by using a generalized linear regression model to deal with different types of data [13,22]. Different probabilistic link functions are used to establish the regression between observed data and low-dimension latent variables. The method also assumed sparse relationships between observed data and latent variables using l_1-norm penalty in the regression model. Low-dimension latent model is directly related with non-negative matrix factorization (NMF). Zhang *et al*. proposed a joint NMF method to find the shared feature matrix across multi-omics datasets, which is a low-dimension representation of the original high-dimension data [16,23]. After transforming multi-omics data into the low-dimension subspace, classical clustering methods such as *k*-means can be used to get the final clusters. However, the objective functions of these latent factor models are usually non-convex. They can only use sampling based algorithms to find suboptimal solutions, which are usually slow and unstable. As the quickly increasing data volume, these computational issues will become more serious. As an alternative, low-rank approximation based methods are also used to find the low-dimension subspace of high-dimensional data [24–27], including cancer genomic data [28].We developed a

Table 1. Selected methods for unsupervised clustering of multi-omics data.

Strategy	Description	Methods	Basic model	Refs.
Direct integrative clustering	Put multi-omics datasets into a stacked matrix, and take this stacked matrix as input for the following clustering analysis	Super *k*-means	Direct clustering with BIC	[12]
		iCluster+	Latent factor analysis	[13]
		JIVE	Low rank-based approximation	[14]
		LRAcluster	Low rank-based approximation	[15]
		jNMF	Non-negative matrix factorization	[16]
		Pathifier	Pathway-based integration	[17]
Clustering of clusters	Perform clustering analysis on every single omics dataset and then integrate the primary clustering results into final cluster assignments	COCA	Clustering of intermediated clusters	[3]
		MDI	Latent *Dirichlet* allocation	[18]
		BCC	Latent *Dirichlet* allocation	[19]
		SNF	Similarity network fusion	[20]
Regulatory integrative clustering	Focus on driver variations by considering the regulatory structures between different molecular layers	PARADIGM	Pathway-based integration	[21]

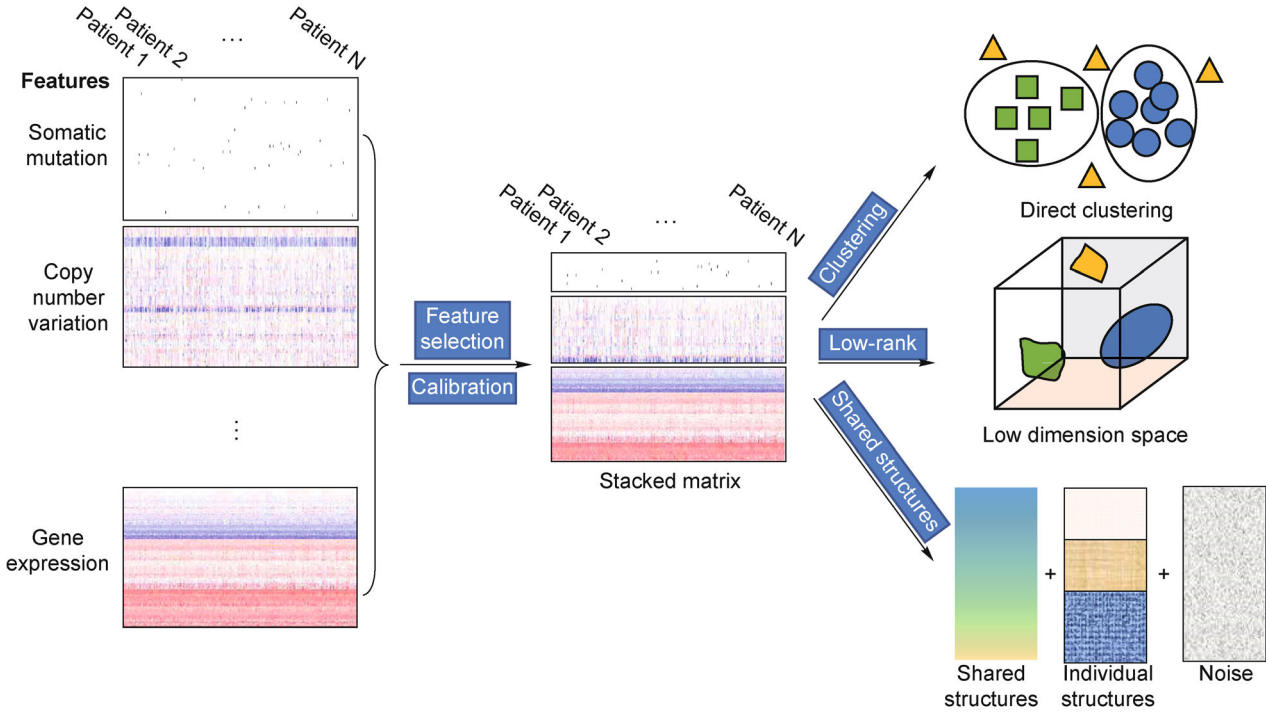

Figure 1. Direct integrative clustering. This kind of methods first stacks multi-omics datasets as a single matrix, and takes this stacked matrix as input to the subsequent clustering analysis.

low-rank approximation based integrative clustering method LRAcluster, which can quickly and stably find the global optimal with simple gradient-ascent algorithm [15].

The second strategy can be extended to an advanced joint and individual variation model, which assumes that different omics datasets share a joint variation pattern and each dataset also has another addictive individualized variation pattern. Joint and Individual Variation Explained (JIVE) used this strategy to jointly analyze gene expression and miRNA expression data [14]. Joint clustering across multi-omics data can be done based on the low-rank joint variation matrix, and independent clustering can also be implemented using the individual variation matrix (also with low-rank assumption). But this method can only deal with Gaussian data. The combination of above generalized linear regression model with JIVE model can solve this problem. One major drawback of latent variable or low-rank approximation model is that it is still hard to interpret the biological meaning of the latent factors, although several statistical methods have been proposed to find the associated molecular features with the latent factors [29,30]. Another problem is that the explained percentage of the total variances is frequently small in the low-dimension subspace, which questions the basic assumption that a small number of driving oncogenic processes dominate the molecular variation patterns in cancer.

Assembling multi-omics data into one single matrix is the core consideration of direct integrative clustering. After the transformation, most clustering methods, such as *k*-means, hierarchical clustering, as well as other density-based and network-based methods [31–34] can be used to generate different views of molecule-based cancer classifications. For the second strategy, the molecular features associated with different dimensions in the subspace can be used to analyze enriched biological pathways and functions. This information is important for biologically interpreting the major variation patterns (it is commonly assumed that each dimension in the subspace is related to a separate oncogenic process [13,16]) and the identified clusters. One drawback of direct integrative clustering is that if the dimensions orvariances of different omics datasets differ a lot, the integrative clustering results may be biased to those datasets with larger dimension or larger variance. Necessary data pre-processing procedure, which will be discussed in "Method in Practice" section, should be made before the analysis.

Clustering of clusters

Unlike direct integrative clustering which implements clustering by directly stacking different types of multi-omics data, clustering of clusters (COC) is another way which performs clustering analysis on every single omics dataset and then integrates the primary clustering results

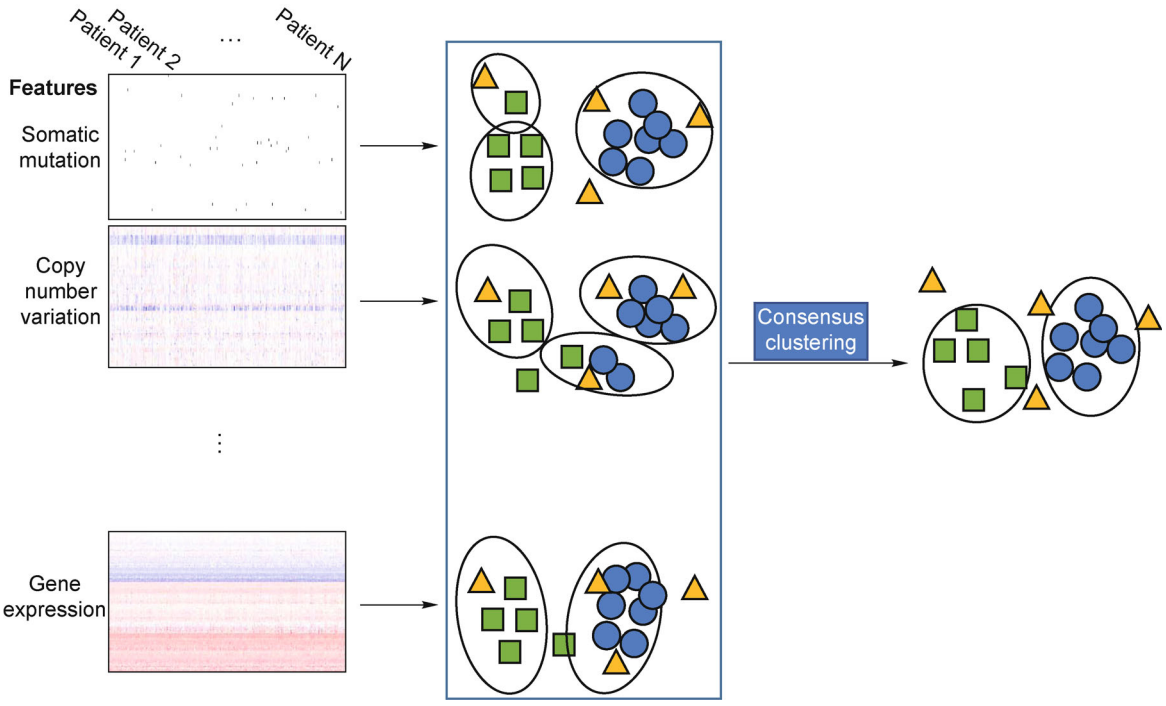

Figure 2. Clustering of clusters. This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.

into final cluster assignments (Figure 2). Primary clustering results can be regarded as an intermediate representation of dataset-specific similarities. And final clustering assignments are expected to make an appropriate comprise based on these similarities.

A direct strategy is to code the primary clustering results of each omics dataset into a binary vector which indicates the sample-cluster assignments. For example, if we perform *k*-means clustering on gene expression data and obtain three clusters, samples belonging to the first cluster would be encoded by [1,0,0], the second cluster [0,1,0] and the third [0,0,1]. Repeating this process for every dataset, we can code every sample into a new binary vector whose length is $\sum_{i=1}^{t} k_i$ where t is the number of datasets and k_i is the number of clusters for dataset i. This vector forms a concise representation of original omics data variations in terms of similarity between samples. Based on the new feature vectors, subsequent clustering analysis can be implemented to get overall cluster assignments [3,8].

Instead of "hard" sample-cluster assignments, some probabilistic models can generate "soft" or probabilistic assignments, such as Gaussian mixture models or *Dirichlet* allocation process [35–37]. This kind of method usually describes a generative process, which models the cluster label as a hidden variable, controlling the probabilistic distributions that "generate" observed omics data. For dealing with multiple datasets, two layers of hidden variables are needed to represent both the overall clustering assignments and dataset-specific clustering assignments. A method named Bayesian consensus clustering was developed based on this strategy [19]. In its three-layer Bayesian network model, they assumed for every single sample of m different omics types $x_i = (x_{ij}, j = 1, \ldots, m)$, there exists an overall clustering assignment C_i which follows a multinomial distribution with a *Dirichlet* conjugate prior.Then the separate clustering result L_{ij} for j-th dataset of this sample adheres loosely to the overall cluster by a function $v(C_i, L_{ij})$, which is inclined to favor the consistence between different datasets. Finally, the observation genomic data x_{ij} is drawn from a mixture distribution specified by L_{ij}. By the sematic of Bayesian networks, x_{ij} is independent of C_i, given L_{ij}. Through effective learning and inference algorithm (such as Gibbs sampling), the method can obtain the probability of sample-cluster assignment C_i for every sample. Another example applying similar *Dirichlet* allocation model, is called MDI (multiple datasets integration) [18,38]. The general method of integrating multiple separate clustering results into one overall assignment is called consensus clustering, which has been studied well in the field of machine learning. They could be considered in different contexts [39,40].

Above methods need first to learn the sample-cluster assignments in each omics dataset. Another strategy tries to directly combine the sample similarities in each dataset to an integrated similarity network. Then, network-based clustering methods, such as community-structure analysis [41,42], spectral clustering [43,44] and Markov clustering algorithm (MCL) [45], can be used to identify the final clusters. A method called SNF (similarity network fusion) first built a patients' similarity network for each dataset, based on the 'distance' of each pair of patients, and then it used a message passing algorithm to combine the similarities from multiple datasets [20]. In its original paper, the method only analyzed mRNA, miRNA expression and DNA methylation data, which are all real number data and can be measured by Euclidean distance. Other distance measures should be tested for analyzing other kinds of omics data.

The main consideration of COC is whether the intermediate structures or similarities can maintain the clustering information in the original datasets. And the transformation will make it more difficult to detect shared structures among different omics datasets. Another problem is that if there is a great disparity between the clustering results of different omics datasets, it should be checked whether the following consensus clustering is meaningful or not. Although this problem exists almost for all kinds of integrative methods, it is more serious for COC. Some measures should be applied to evaluate the consistence of different clustering results, such as Jaccard similarity [46], adjusted Rand index [47] and other information theory-based measures [33].

Regulatory integrative clustering

Besides cancer driver molecular alterations which play essential roles in cancer initiation and progression, there are also many passenger alterations co-occurring at the same time. To reduce the "noises" caused by passenger alterations, molecule-based cancer classifications should focus on driver alterations. The basic idea of regulatory integrative clustering is to only use the driver variations by considering the regulatory structures between different molecular layers, which can help discriminate driver and passenger variations (Figure 3).

Candidate cancer drivers are usually defined as significant genetic alterations, such as somatic mutations and copy number variations [48–51]. Recurrent epigenetic alterations, especially DNA methylation [52], and large-scale functional screens [53–56] can also be used to identified novel candidate drivers. Then, these alterations or perturbations are integrative analyzed with gene expression variations (either *cis*- or *trans*-acting effects), the functional indications of genetic alterations [57–61]. TCGA researchers have used this method to identify molecular subtypes of various cancers, such as the $BRAF^{V600E}$-like and RAS-like subtypes of papillary thyroid carcinoma [62].

Compared with direct integrative clustering and clustering of clusters, the clusters identified by regulatory integrative clustering methods are clearer on their molecular mechanisms. This kind of method is more like a framework or pipeline for integrated analysis instead of a general algorithm.

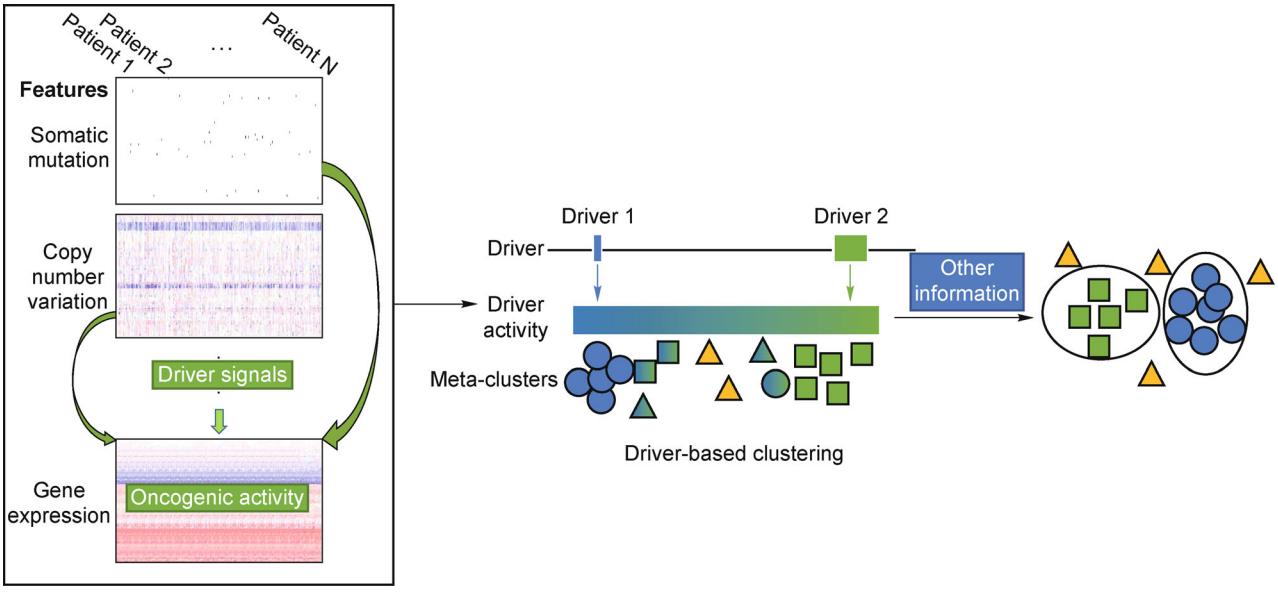

Figure 3. Regulatory integrative clustering. This kind of methods only uses driver variations by considering the regulatory structures between different molecular layers.

METHOD IN PRACTICE

Pre-processing and post-processing

Data pre-processing is crucial for multi-omics integrative analyses in terms of reducing unwanted biases and noises. Improper pre-processing frequently causes misleading results. Major pre-processing steps include: feature selection, data normalization and dimension reduction:

i) Feature selection is to remove irrelevant molecular features, which have small variations or have poor correlations with clinical outputs, such as survivals and drug responses. Variance analysis, correlation analysis and some advanced Bayesian strategies can be used to remove the features with small variances. And, bootstrap based methods such as random forest or sparse regression models can be used to select clinical relevant features.

ii) Normalization is to reduce non-biological batch effects [63] and data variations. Quantile normalization and invariant set normalization [64] are most commonly used to reduce batch effects.

iii) Dimension reduction is to extract the major variations of high-dimensional data. It can greatly reduce data complexity and computational costs. Some popular unsupervised dimension reduction methods are broadly used, such as principal component analysis (PCA), singular value decomposition (SVD), t-distributed stochastic neighbor embedding (t-SNE) [65] and non-negative matrix factorization (NMF) [66,67]. But most of these methods assume Gaussian distribution of data. When applied to other data types, they should be modified accordingly.

In practice, a study may not take all the pre-processing steps and a few methods can directly include some steps in the computational models.

It is important for biologically interpreting the clustering results to perform post-clustering analyses, routinely including molecular signature analysis and clinical outcome analysis:

(i) Signature or differential analysis is to identify the molecular features associated with each cluster. For example, differentially expressed genes can be identified by comparing gene expressions in each cluster against the other clusters. Then, enriched pathways and gene ontologies of these genes can be used to annotate the biological characteristics of each cluster.

(ii) Clinical outcome analysis is important to evaluate the identified clusters. Survival analysis is commonly implemented to evaluate the prognostic value of the molecule-based subtypes. Also, some studies try to combine clinical variables with identified molecular subtypes for better or more precise patient stratification.

Necessary refinements of the clustering results should be made according to the post-clustering analyses.

Pathway and network based clustering

Pathways and molecular networks contain useful prior knowledge of biological processes and functions. Existing databases have collected many pathways and general gene networks from biomedical literatures and high-throughput experiments [68–70]. Many studies suggest that incorporation of this prior information can improve the performances of computational analysis models. Pathway and gene network information are also widely used to cluster multi-omics data for molecule-based cancer classifications.

The simple strategy is to map all the omics datasets to annotated pathways or networks regardless of their molecular layers. Pathifier proposed a "principal curve" distance based pathway-level score, which is calculated based on the omics data of all the pathway genes, to describe the activity of each pathway in each sample [17,71]. For each pathway analyzed, Pathifier supposes a d_p-dim pathway subspace of the pathway genes (d_p is the number of genes in the pathway) and every sample can be represented by a point in this space. Then, a "principal curve" is fitted to all sample points in the subspace. The distance of the sample point to the fitted curve is used as the pathway activity score for each sample. In its original study, only gene expression data were used, but this method can be easily extend to other types of omics data. Other methods to obtain alterations of pathways include, SPIA [72], TieDIE [73], and so on.

Some advanced approaches try to model inter-layer regulatory structures to score the pathway level alterations. PARADIGM used a directed factor graph (Bayesian network) to model various inter-/intra-layer interactions including transcription, translation, protein activation, protein complex formation and gene family redundancy for any given NCI pathway [21]. The variables in the graph represent different biological entities, such as copy numbers, mRNA expressions, protein expressions and protein activities in the original model. And the directed edges represent regulatory or signaling effects between two variables. The global likelihood of the Bayesian network based on the observed multi-omics data was calculated as the pathway level score for each sample. However, the major disadvantage of this method is that the factor graph model strongly relies on the prior knowledge of annotated pathways. Missing links or cell-type specific effects may significantly change the results.

Instead of pathway-level aggregation, another strategy uses network information to smooth omics data. Hofree *et al*. proposed a network-based stratification (NBS) to integrate mutation data with gene networks [74]. The

NBS model assumed that a gene mutation may not only *cis*-regulate its own but also affect its interacted genes in the network. Liu and Zhang adopted this model to jointly analyze genomic and epigenetic alterations (including somatic mutations, copy number variations and DNA methylations) with a pre-given gene network [75]. The gene network is used to smooth the variations in different omics datasets. Then, joint non-negative matrix factorization (jNMF) was used to find the pan-cancer subgroups [16].

Although there are still many noises in pathway and gene network annotations, such as missing links, cell-type specific effects and parameter dynamics, this kind of prior knowledge is very useful for establishing more biological meaningful computational models.

Applications

TCGA provides a unified platform for practical applications of multi-omics unsupervised clustering analyses. TCGA pan-cancer study intends to provide another cancer classification system based on molecular features rather than their tissue origins [76]. By integrative Cluster-of-Cluster Assignments (COCA) analysis of somatic mutation, copy number variation, DNA methylation, mRNA expression, miRNA expression and protein activity data across 3,527 samples from 12 different cancer types, they identified 11 molecule-based major subtypes [3]. Five subtypes are nearly identical to the tissue-of-origin classification, but the remaining samples are re-grouped, for example, lung squamous carcinomas, head/neck squamous carcinomas and some of bladder carcinomas are grouped as a separate squamous-like subtype [3]. Several other methods, such as LRAcluster [15] and network-based jNMF [75], have also been applied for pan-cancer classifications. They also found a cross tissue origin squamous-like subtype. In TCGA gastric cancer analysis, they used two different methods, iCluster+ [13] and clustering of subtype assignments, to analyze multi-omics data. Then, the clustering results were combined as the final four major subtypes [77]. To obtain more biologically meaningful results, they analyzed various features' significance in each single cluster and then made subtle adjustments of original clustering results.

In real applications, there are much more details that need to be considered. Different study designs or different cancer types may favor different clustering methods. As the release of more large-scale cancer multi-omics datasets, the advantages and disadvantages of different integrative clustering methods can be systematically evaluated.

DISCUSSION

Unsupervised clustering analysis of multi-omics data is an important way to discover novel molecule-based cancer subtypes. In this review, we generally summarized current methods into three major categories, direct integrative clustering, clustering of clusters and regulatory integrative clustering, based on their different strategies on dealing with the data type heterogeneity and inter-omics regulatory structures. Direct integrative clustering is straightforward, but effective in most real cases, by stacking multiple datasets into a single matrix. One major approach, the joint subspace based integrative clustering can also find the shared principal subspaces and the associated molecular features across different omics datasets. Clustering of clusters solves the problem of data heterogeneity by transforming every single omics dataset into an intermediate form, such as sample-cluster assignments or sample similarity networks. But this strategy is hard to find shared data structures, and most data variation information is lost during the transformation. The third strategy, regulatory integrative clustering takes more inter-omics regulatory structures into computational models. In practices, pathway and gene networks are commonly used to reduce the complexity of computational models.

Currently, one aim of precision oncology is to identify new molecule-based cancer subtypes from large-scale cancer multi-omics data. However, some other information should also be considered. For example, clinical features, such as ages, genders, pathological stages and blood test records, can better predict the overall survivals than omics data [78]. Unlike the omics data which only contain molecular level information of cancer tissues, clinical features provide more macro-scale perspectives of patients. Some clinical guidelines of targeted therapies contain both clinical and genomic indications. Unsupervised methods to integrate both clinical features and multi-omics data are still lacking. The principle of partial least squares (PLS) [79,80] may be used for developing this kind of methods. Another important source is functional genomics data. High-content screening techniques using RNAi [53] or CRISPR/Cas9 [55]generated many genome-scale genetic interaction and gene-phenotype association data. These data can help describe causal mechanisms between genetic variations and gene activity, and then help obtain new subtypes. It is promising and challenging for integrating omics data with other kinds of biomedical data for more precise cancer classifications in near future.

As the decreasing cost of high throughput omics techniques, cancer omics data will increase much faster in following decades. More effective infrastructure for data

storage and sharing should be developed [81]. For integrative clustering, the methods should take more considerations on the increasing computational burden in future, such as memory requirement, parallel computing ability, and especially time cost.

ACKNOWLEDGEMENTS

This work is supported by National Basic Research Program of China (No. 2012CB316503), National Natural Science Foundation of China (Nos. 61370035 and 31361163004) and Tsinghua University Initiative Scientific Research Program.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Dongfang Wang and Jin Gu declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Garraway, L. A., Verweij, J. and Ballman, K. V. (2013) Precision oncology: an overview. J. Clin. Oncol., 31, 1803–1805
2. Shrager, J. and Tenenbaum, J. M. (2014) Rapid learning for precision oncology. Nat. Rev. Clin. Oncol., 11, 109–118
3. Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D., Niu, B., McLellan, M. D., Uzunangelov, V., *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 158, 929–944
4. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. and Kim, D. (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat. Rev. Genet., 16, 85–97
5. Liu, Z., Zhang, X. S. and Zhang, S. (2014) Breast tumor subgroups reveal diverse clinical prognostic power. Sci. Rep., 4, 4002
6. Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., Diao, L., Xu, Y., Verhaak, R. G. and Liang, H. (2014) The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat. Commun., 5, 3963
7. 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, 346–352
8. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. Nature, 490, 61–70
9. Popat, S., Hubner, R. and Houlston, R. S. (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. J. Clin. Oncol., 23, 609–618
10. Issa, J. P. (2004) CpG island methylator phenotype in cancer. Nat. Rev. Cancer, 4, 988–993
11. Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K., Frigessi, A. and Børresen-Dale, A. L. (2014) Principles and methods of integrative genomic analyses in cancer. Nat. Rev. Cancer, 14, 299–313
12. Zhang, W., Liu, Y., Sun, N., Wang, D., Boyd-Kirkup, J., Dou, X. and Han, J. D. (2013) Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Reports, 4, 542–553
13. Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M. and Shen, R. (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. USA, 110, 4245–4250
14. Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013) Joint and Individual Variation Explained (Jive) for integrated analysis of multiple data types. Ann. Appl. Stat., 7, 523–542
15. Wu, D., Wang, D., Gu, J. and Zhang, M. Q. (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genomics, 16, 1022.
16. Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W. and Zhou, X. J. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res., 40, 9379–9391
17. Drier, Y., Sheffer, M. and Domany, E. (2013) Pathway-based personalized analysis of cancer. Proc. Natl. Acad. Sci. USA, 110, 6388–6393
18. Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. and Wild, D. L. (2012) Bayesian correlated clustering to integrate multiple datasets. Bioinformatics, 28, 3290–3297
19. Lock, E. F. and Dunson, D. B. (2013) Bayesian consensus clustering. Bioinformatics, 29, 2610–2616
20. 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. Nat. Methods, 11, 333–337
21. Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J. M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 26, i237–i245
22. Shen, R., Olshen, A. B. and Ladanyi, M. (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
23. Zhang, S., Li, Q., Liu, J. and Zhou, X. J. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics, 27, i401–i409
24. Candes, E. J., Li, X. D., Ma, Y. and Wright, J. (2011) Robust principal component analysis? J. ACM, 58
25. Boyd, S.Parikh, N.Chu, E.Peleato, B.Eckstein,and J.. (2011) Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3, 1–122
26. Candès, E. J. and Recht, B. (2009) Exact matrix completion via convex optimization. Found. Comput. Math., 9, 717–772
27. Cai, J. F., Candes, E. J. and Shen, Z. W. (2010) A singular value thresholding algorithm for matrix completion. SIAM J. Optim., 20, 1956–1982.
28. Zhou, X., Liu, J., Wan, X. and Yu, W. (2014) Piecewise-constant and low-rank approximation for identification of recurrent copy number variations. Bioinformatics, 30, 1943–1949
29. Chung, N. C. and Storey, J. D. (2015) Statistical significance of variables driving systematic variation in high-dimensional data. Bioinformatics, 31, 545–554
30. Linting, M., van Os, B. J. and Meulman, J. J. (2011) Statistical significance of the contribution of variables to the PCA solution: an alternative permutation strategy. Psychometrika, 76, 440–460.
31. Friedman, J., Hastie, T. and Tibshirani, R. (2009) The Elements of Statistical Learning. New York: Springer-Verlag

32. Jain, A. K., Murty, M. N., and Flynn, P. J. (1999) Data clustering: a review. ACM computing surveys (CSUR), 31, 264–323
33. Han, J., Kamber, M. and Pei, J. (2011) Data mining: concepts and techniques: concepts and techniques. San Francisco: Morgan Kaufmann
34. Rodriguez, A. and Laio, A. (2014) Clustering by fast search and find of density peaks. Science, 344, 1492–1496
35. Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003) Latent dirichlet allocation. J. Mach. Learn. Res., 3, 993–1022
36. Nguyen, X. and Gelfand, A. E. (2011) The Dirichlet labeling process for clustering functional data. Stat. Sin., 21, 1249–1289.
37. Dahl, D. B. (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In Bayesian inference for gene expression and proteomics, 201–218, Cambridge: Cambridge University Press
38. Savage, R. S., Ghahramani, Z., Griffin, J. E., Kirk, P. and Wild, D. L. (2013) Identifying cancer subtypes in glioblastoma by combining genomic, transcriptomic and epigenomic data. arXiv:1304.3577
39. Nguyen, N. and Caruana, R. (2007) Consensus clusterings. In Data Mining, ICDM 2007. Seventh IEEE International Conference, 607–612
40. Goder, A. and Filkov, V. (2008) Consensus Clustering Algorithms: Comparison and Refinement. in Alenex, SIAM., 109–117
41. Girvan, M. and Newman, M. E. (2002) Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 99, 7821–7826
42. Newman, M. E. (2006) Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA, 103, 8577–8582
43. Ng, A. Y., Jordan, M. I. and Weiss, Y. (2001) On spectral clustering: Analysis and an algorithm. In Advances in Neural Information Processing Systems. 849–856, Cambridge: MIT Press
44. von Luxburg, U. (2007) A tutorial on spectral clustering. Stat. Comput., 17, 395–416.
45. Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res., 30, 1575–1584
46. Levandowsky, M. and Winter, D. (1971) Distance between sets. Nature, 234, 34–35.
47. Hubert, L. and Arabie, P. (1985) Comparing partitions. J. Classif., 2, 193–218.
48. Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., *et al.* (2015) Toward understanding and exploiting tumor heterogeneity. Nat. Med., 21, 846–853
49. Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., Yue, P., Haverty, P. M., Bourgon, R., Zheng, J., *et al.* (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature, 466, 869–873
50. Lohr, J. G., Stojanov, P., Lawrence, M. S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y. W., Slager, S. L., *et al.* (2012) Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc. Natl. Acad. Sci. USA, 109, 3879–3884
51. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature, 499, 214–218
52. Villanueva, A., Portela, A., Sayols, S., Battiston, C., Hoshida, Y., Méndez-González, J., Imbeaud, S., Letouzé, E., Hernandez-Gea, V., Cornella, H., *et al.* (2015) DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. Hepatology, 61, 1945–1956
53. Eifert, C. and Powers, R. S. (2012) From cancer genomes to oncogenic drivers, tumour dependencies and therapeutic targets. Nat. Rev. Cancer, 12, 572–578
54. Sanchez-Garcia, F., Villagrasa, P., Matsui, J., Kotliar, D., Castro, V., Akavia, U. D., Chen, B. J., Saucedo-Cuevas, L., Rodriguez Barrueco, R., Llobet-Navas, D., *et al.* (2014) Integration of genomic data enables selective discovery of breast cancer drivers. Cell, 159, 1461–1475
55. Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. Science, 343, 84–87
56. Jiang, P., Wang, H., Li, W., Zang, C., Li, B., Wong, Y. J., Meyer, C., Liu, J. S., Aster, J. C. and Liu, X. S. (2015) Network analysis of gene essentiality in functional genomics experiments. Genome Biol., 16, 239
57. Chen, J. C., Alvarez, M. J., Talos, F., Dhruv, H., Rieckhof, G. E., Iyer, A., Diefes, K. L., Aldape, K., Berens, M., Shen, M. M., *et al.* (2014) Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. Cell, 159, 402–414
58. Fehrmann, R. S., Karjalainen, J. M., Krajewska, M., Westra, H. J., Maloney, D., Simeonov, A., Pers, T. H., Hirschhorn, J. N., Jansen, R. C., Schultes, E. A., *et al.* (2015) Gene expression analysis identifies global gene dosage sensitivity in cancer. Nat. Genet., 47, 115–125
59. Rockman, M. V. and Kruglyak, L. (2006) Genetics of global gene expression. Nat. Rev. Genet., 7, 862–872
60. Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A. and Pe'er, D. (2010) An integrated approach to uncover drivers of cancer. Cell, 143, 1005–1017
61. Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S. and Freedman, M. L. (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell, 152, 633–641
62. Cancer Genome Atlas Research Network. (2014) Integrated genomic characterization of papillary thyroid carcinoma. Cell, 159, 676–690
63. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. and Irizarry, R. A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet., 11, 733–739
64. Eisenberg, E. and Levanon, E. Y. (2003) Human housekeeping genes are compact. Trends Genet., 19, 362–365
65. van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. J. Mach. Learn. Res., 9, 2579–2605.
66. Hoyer, P. O. (2004) Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res., 5, 1457–1469.
67. Lee, D. D. and Seung, H. S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature, 401, 788–791
68. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res., 40, D109–D114
69. Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res., 39, D691–D697
70. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res.,

42, D459–D471
71. Livshits, A., Git, A., Fuks, G., Caldas, C. and Domany, E. (2015) Pathway-based personalized analysis of breast cancer expression data. Mol. Oncol., 9, 1471–1483
72. Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., Kim, C. J., Kusanovic, J. P. and Romero, R. (2009) A novel signaling pathway impact analysis. Bioinformatics, 25, 75–82
73. Paull, E. O., Carlin, D. E., Niepel, M., Sorger, P. K., Haussler, D. and Stuart, J. M. (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics, 29, 2757–2764
74. Hofree, M., Shen, J. P., Carter, H., Gross, A. and Ideker, T. (2013) Network-based stratification of tumor mutations. Nat. Methods, 10, 1108–1115
75. Liu, Z. and Zhang, S. (2015) Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genomics, 16, 503
76. Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. andStuart, J. M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet., 45, 1113–1120
77. Cancer Genome Atlas Research Network. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 513, 202–209
78. Yuan, Y., Van Allen, E. M., Omberg, L., Wagle, N., Amin-Mansour, A., Sokolov, A., Byers, L. A., Xu, Y., Hess, K. R., Diao, L., *et al.* (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol., 32, 644–652
79. Wold, S., Martens, H. and Wold, H. (1983) The multivariate calibration-problem in chemistry solved by the Pls Method. Lect. Notes Math., 973, 286–293.
80. Bastien, P., Bertrand, F., Meyer, N. and Maumy-Bertrand, M. (2015) Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics, 31, 397–404
81. Aronson, S. J. and Rehm, H. L. (2015) Building the foundation for genomics in precision medicine. Nature, 526, 336–342

