W154–W159 *Nucleic Acids Research, 2016, Vol. 44, Web Server issue* *Published online 30 April 2016*
doi: 10.1093/nar/gkw378

ICM: a web server for integrated clustering of multi-dimensional biomedical data

Song He†, Haochen He†, Wenjian Xu, Xin Huang, Shuai Jiang, Fei Li*, Fuchu He* and Xiaochen Bo*

Beijing Institute of Radiation Medicine, Beijing 100850, P. R. China

Received January 31, 2016; Revised April 24, 2016; Accepted April 25, 2016

ABSTRACT

Large-scale efforts for parallel acquisition of multi-omics profiling continue to generate extensive amounts of multi-dimensional biomedical data. Thus, integrated clustering of multiple types of omics data is essential for developing individual-based treatments and precision medicine. However, while rapid progress has been made, methods for integrated clustering are lacking an intuitive web interface that facilitates the biomedical researchers without sufficient programming skills. Here, we present a web tool, named Integrated Clustering of Multi-dimensional biomedical data (ICM), that provides an interface from which to fuse, cluster and visualize multi-dimensional biomedical data and knowledge. With ICM, users can explore the heterogeneity of a disease or a biological process by identifying subgroups of patients. The results obtained can then be interactively modified by using an intuitive user interface. Researchers can also exchange the results from ICM with collaborators via a web link containing a Project ID number that will directly pull up the analysis results being shared. ICM also support incremental clustering that allows users to add new sample data into the data of a previous study to obtain a clustering result. Currently, the ICM web server is available with no login requirement and at no cost at http://biotech.bmi.ac.cn/icm/.

INTRODUCTION

With the rapid development of high-throughput technologies, parallel acquisition of multiple types of omics data for a disease or a bioprocess is becoming less expensive. Annotations for genes, proteins and drugs are also growing rapidly. The construction of large-scale repositories of multi-dimensional biomedical data is underway. For example, the International Cancer Genome Consortium (1), The Cancer Genome Atlas (TCGA) (2) and the Cancer Genome Project (3) have already accumulated multi-dimensional biomedical data for cancer patients, including genomics, transcriptomics, proteomics and epigenomics data. As a result, researchers can now explore the heterogeneity of a disease or a biological process by examining multiple types of data to obtain a comprehensive view (4–6). To achieve this, methods and software for multi-omics studies, especially integrated clustering analysis (7–10), have become valuable resources for researchers. Furthermore, the integrated clustering of multi-dimensional biomedical data is particularly important for various precision medicine projects whose aims include the identification of novel therapeutic schedules based on an extensive characterization of biologic specimens (11).

There are several methods that have been used for the integration of multi-dimensional biomedical data. 'Concatenation' is a commonly used method that is simple and has a low computational-cost. With this method, each sample with multi-dimensional features can be assembled into a long integrated vector that maintains the complete information profile of the sample. Conversely, the 'iCluster' (7,10) method which is based on a Gaussian latent variable model effectively discovers potentially novel subclasses from multi-dimensional data, while potentially excluding certain features in order to reduce the number of calculations needed for the processing of multi-dimensional data. To address computational complexity without the preferential accommodation of certain features over others, Wang *et al*. (8) presented the 'SNF' method which can explore communities of specimens with an integrated multi-dimensional network. This network represents a fusion of similar networks of samples based on a single type of data respectively.

Several programing packages have been developed for the methods mentioned above, and these require users to be familiar with programming languages and additional bioinformatics tools. Therefore, web tools with interactive

*To whom correspondence should be addressed. Tel: +86 10 66931207; Fax: +86 10 66931242; Email: boxiaoc@163.com
Correspondence may also be addressed to Fuchu He. Tel: +86 10 68171208; Fax: +86 10 80705155; Email: hefc@bmi.ac.cn
Correspondence may also be addressed to Fei Li. Tel: +86 10 66931422; Fax: +86 10 66931242; Email: pittacus@gmail.com
† These authors contributed equally to the paper as first authors.

analysis interfaces and intuitive visualization are needed for most biomedical researchers that do not have sufficient programing skills. Furthermore, web applications with result-sharing mechanisms are also becoming more important, especially for large-scale cooperative research projects that depend on an analysis of specimens that have different collection sources.

To address the growing demand for exploring multi-dimensional biomedical data for novel insights, we have developed a web server named ICM (Integrated Clustering of Multi-dimensional biomedical data), which provides analysis tools for the fusion, clustering and visualization of multi-dimensional biological data and knowledge. ICM is currently available at http://biotech.bmi.ac.cn/icm/ and has no login requirement or cost. ICM employs three typical algorithms, 'SNF', 'iCluster' and 'Concatenation', so that users can explore the heterogeneity of a disease or a biological process based on available multi-dimensional data. For example, subtypes of cancer can be identified by performing an integrated clustering of gene expression, DNA methylation and miRNA expression data (6). The clustering results obtained can then be arranged, illustrated and modified by ICM with an intuitive user interface. In addition, ICM supports remote cooperation via assignment of Project ID numbers that are associated with each analysis result. For efficiency, ICM also supports incremental clustering by which users can obtain clustering results after adding new sample data to their previous study that was assigned a Project ID number.

MATERIALS AND METHODS

Implementation

The ICM framework has two main components. One is an interface that implements interactive analysis and result visualization, while the second component includes the computational services provided by the server. The interface of ICM was written using the Bootstrap framework for HTML, CSS and JavaScript. 'Datatables' (a plug-in of jQuery) are used to show tabular data and 'Cytoscape web' (a plug-in of flash) is used to show an interactive clustering network. The D3 library of JavaScript is also used to illustrate heatmaps. Concerning the load capacity and response time of the browser, static pictures are used to visualize data when the latter becomes too large. For the computational services, the spring 'mvc' framework for the Java program language was used. Finally, storage and management of the application data are implemented by MySQL.

Data preprocessing

Prior to data analysis, users can edit the name and type (continuous or discrete) of their data, and then selectively apply log2 transformation and standardization. Similarities between samples are measured by pairwise Euclidean distances or pairwise Chi-squared distances for continuous or discrete data, respectively. If users choose to standardize their data, each row of the data is standardized with an average of 0 and a standard deviation of 1.

Data sets of application case

For a case study of ICM, DNA methylation, mRNA sequence and miRNA sequence data for 72 patients with acute myeloid leukemia (LAML) were obtained from TCGA (2) (Firehose run of the Broad Genome Data Analysis Center [http://gdac.broadinstitute.org]). The clinical data of these patients were also downloaded. Before uploading these data to ICM, data cleaning was performed to improve data set quality using the following steps:

(i) If more than 20% of the data for a patient across the features examined were missing, the patient data were filtered out. Similarly, if there were more than 20% of the data for a particular feature that was missing across the patients examined, that feature data were filtered out.
(ii) Average imputations were used to fill in missing data.
(iii) Each row (feature) of the data was standardized with an average value of 0 and a standard deviation of 1.

RESULTS

Overview of the user interface

ICM is able to be accessed by most web browsers, including Google Chrome, Mozilla Firefox, Safari and Internet Explorer (v11+). Google Chrome is recommended to achieve the best visualization.

After an analysis method is selected in the 'Analysis' icon in ICM, users should upload at least two data files including feature-by-sample matrix. A preview of the uploaded data is available in the webpage. By setting parameters and running the analysis application selected, users can interactively explore the results in network graphs and tables that are provided as output on the webpages. For the 'SNF' and 'Concatenation' methods, the important features for the clustering results are also identified and illustrated in the 'Feature' subpage (Figure 1A). All of the results can be downloaded and accessed via a unique project ID number and web link for 30 days. Users can also share the results obtained with collaborators by providing the unique project ID number that was assigned (Figure 1B). Default settings are available for each step, and on the first page of each method, the 'Start your tour' icon can guide users through each of the necessary steps.

Data input

Each method in ICM has the same requirement that users should upload more than two data files including feature-by-sample matrix. The rows in the data files represent biological features (e.g. gene symbol, Uniprot ID, miRNA symbol, etc.), while the columns represent samples (e.g. clinical samples, cellular samples, etc.). Moreover, the first column and row of the matrix should include sample names and feature names, respectively. In each data set, the sample names should be completely consistent with those of the remaining data sets. Except for feature and sample names, all data should be numeric. Furthermore, the data files should be in '.txt', '.xlsx', '.xls', '.csv' or '.tsv' formats. After uploading the data sets, users can rename each separately; and

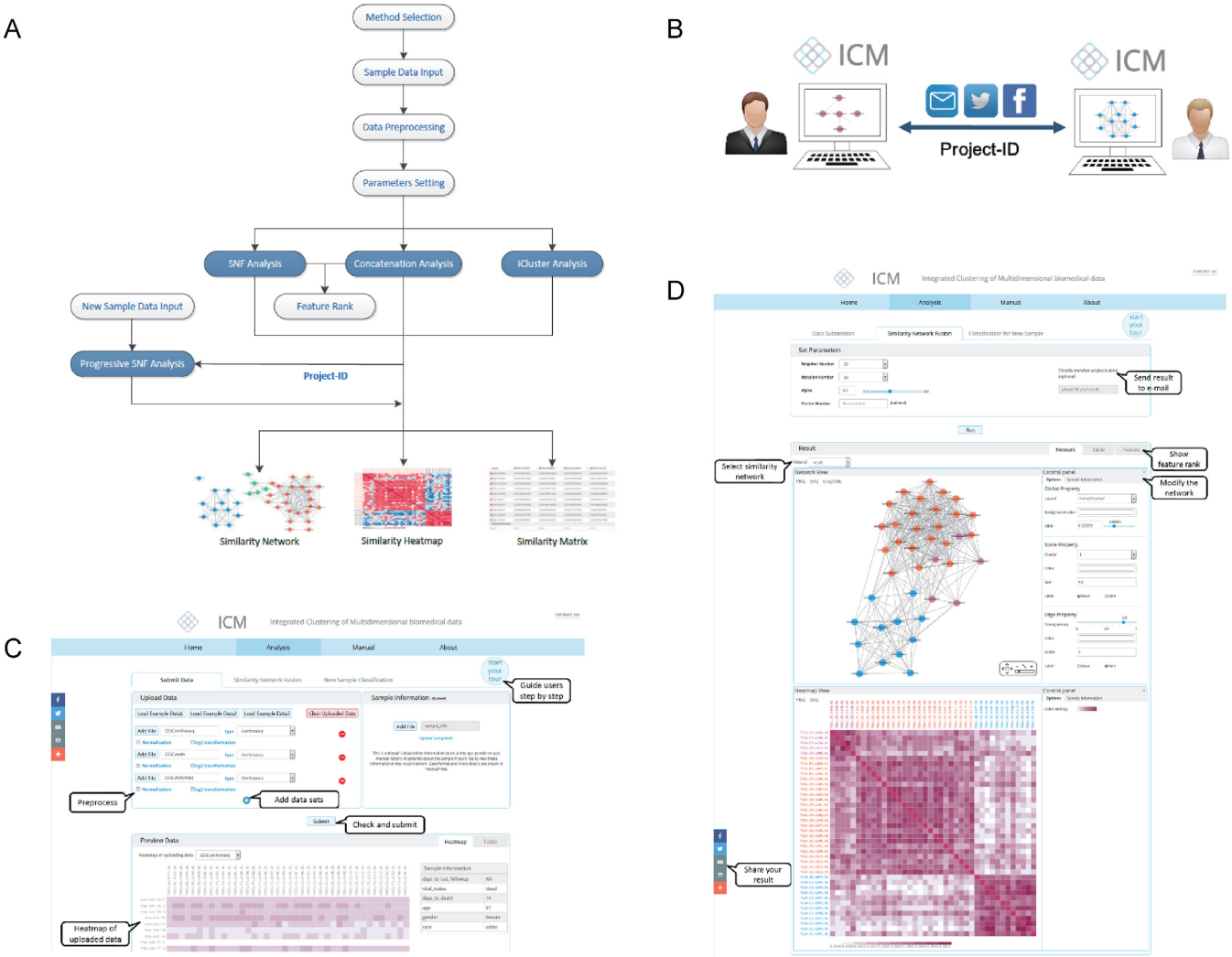

Figure 1. Overview of the ICM framework and its output. (**A**) Overview of the ICM framework. (**B**) Overview of the remote cooperation that is facilitated by ICM. Accordingly, users can exchange results with collaborators by referencing Project ID numbers. (**C**) A snapshot of the ICM start page. The input data field is on the top and the preview data are on the bottom. The 'Start your tour' icon can guide a user through each step. (**D**) A snapshot of an ICM result page. The parameter settings field is on the top, while the similarity network and heatmap for a set of patients are shown on the bottom. Furthermore, the latter can be modified interactively.

the names of the result similarity networks will be consistent with these names. The default name of each data set is the filename of the corresponding data (Figure 1C).

Users can also upload other information for samples (such as clinical data) and then interactively view all of the information provided for the samples in the 'Preview Data' and 'Result' subpages. This is optional. There will be a column for each sample, while each row will represent a particular property of the samples (e.g. gender, age, past medical history). The format of the information files should be '.txt', '.xlsx', '.xls', '.csv' or '.tsv'.

Prior to submission, users can selectively preprocess the input data in the webpage and apply standardization or a log2 transformation. In the 'Preview Data' page, users can then preview and delete the input data in real time. Concerning the load capacity and response time of the browser, users can view a static heatmap if the number of uploaded data points is greater than 100 000 (e.g. sample number * feature number > 100 000). For iCluster method, due to the high computational complexity for high dimension data, users may preselect features in order to ensure that the number of features is less than 2000 (12,13).

Parameter settings

There are different parameters for each analysis method in ICM. After clicking the 'submit' icon in the 'Submit data' page, users can enter the parameters setting page. For the 'SNF' method, there are four parameters: 'neighbor number', 'iteration number', 'alpha' and 'cluster number'. For the 'Concatenation' method, there are three parameters: 'neighbor number', 'alpha' and 'cluster number'. For the 'iCluster' method, there are two parameters: 'penalty term' and 'cluster number'. Users can check the meaning of each parameter in the 'Manual' page of ICM.

Results page

For the 'SNF' method and 'Concatenation' method, the result page includes three subpages: 'Network', 'Table' and 'Feature' (Figure 1D), while 'iCluster' method includes only two subpages, 'Network' and 'Table'. The network graph and heatmap of the similarity network of samples are illustrated in the 'Network' subpage. The numerical similarity matrix of the samples is shown in the 'Table' subpage. In the 'Feature' subpage, users can view the importance rank assigned to the features of each data set that is submitted for integrated clustering. For the 'SNF' method, a separate similarity network for each uploaded data set can be viewed by selecting that data set in the drop down menu in the top left corner of the 'Network' and 'Table' subpages. While a single similarity network is generated for the 'Concatenation' and 'iCluster' method.

Nodes in the similarity network represent samples, while the edges of the network characterize the similarity between samples. Nodes that have the same color represent a cluster of samples. Users can interactively modify the network graphs as follows:

(i) Set a threshold to filter out of the edge that is lower than the threshold. The default threshold will remove half of the edges.
(ii) Set the layout of the similarity network.
(iii) Set the background color of the similarity network.
(iv) Set the color and size of nodes in the same cluster.
(v) Set the thickness, color and transparency of the edges.
(vi) Choose whether the nodes and edge labels are shown in the similarity network.

Similarity networks are available for download in PNG, SVG and graphML formats. Furthermore, the latter format can be imported into Cytoscape for further analysis.

On the 'Network' subpage, there is also an option to present the similarity network as a heatmap, with users able to set the color shown. The intensity of the color block will indicate extent of similarity between sample A and sample B. The column names and row names will provide the sample names, and the names that have the same color are part of the same samples. The heatmap of the similarity network can be downloaded as PNG and SVG formatted files.

In addition, a sample node in the network can be clicked on to allow users to identify a series of color blocks in the heatmap that are associated with the sample highlighted. Meanwhile, other information of the sample node is illustrated in the right side of 'Network' subpage. When clicking an edge between sample A and sample B in the network, users can see two color blocks in the heatmap which represent the similarity between the two samples.

Users can also search, sort, delete and download the similarity matrix illustrated in the 'Table' subpage.

In each data set, the various features have different importance in the integrated clustering. The importance ranks are estimated using normalized mutual information (NMI). A feature rank table is shown on the left side of the 'Feature' subpage. On the right side of this subpage, a large heatmap of the features that significantly differ according to subtype is available. The features are ranked in the top 1% of all NMI values, and samples are divided into subgroups according to the clustering result. The feature rank table and large heatmap are available for downloading.

Validity and stability of integrated clustering

ICM provides three validity and stability indexes of the integrated clustering included in *clValid* R package (14), which is illustrated in the 'Table' subpage.

The first index, *connectivity*, relates to what extent samples are clustered in the same group as their nearest neighbors. The connectivity index ranges between zero and infinity, and should be minimized (15).

The second index, *Dunn Index*, is the ratio of the smallest distance between samples not in the same group to the largest intra-group distance. The Dunn index ranges between zero and one, and should be maximized (16).

The third index, the average proportion of non-overlap (*APN*), is a measurement of stability of clustering. The APN measures the average proportion of samples not clustered in the same group by clustering based on the complete data and clustering based on the data with 10% features removed. The APN index ranges between zero and one, and should be minimized (17).

Remote collaboration

Any analysis result that is associated with a particular Project ID number is saved on the ICM server, and these results are available at any time via the Project ID number or a web link that corresponds with the Project ID number. This system provides a method by which researchers can share the analysis of their results with collaborators by exchanging the Project-ID or the Project-ID-specific link (Figure 1B). However, due to the limited capacity of our server, analysis results can only be saved for 30 days. If a researcher does want to save their results for a longer period of time, they can contact us to make other arrangements.

Classifying new samples

The 'Classification for New Sample' subpage of the 'SNF' method is specifically designed for clinicians. In this subpage, a user may obtain multiple-omics data (e.g. gene expression, DNA copy number and DNA methylation data) for patients, and then cluster these patients' data into different subgroups for personalized therapy. To facilitate the management of these results, ICM provides progressive classification for this clinical scenario. Users can identify subgroups after adding new sample data and providing the Project ID number of a previous analysis result. The feature and type of added samples should be consistent with those of the previous samples and the new samples will be highlighted in the 'New Sample Classification' page. However, this function is only available for the 'SNF' method.

Case study

To demonstrate the capacity of ICM to analyze data, we collected three types of omics data for 72 patients with LAML from TCGA (2) (Firehose run of the Broad Genome Data

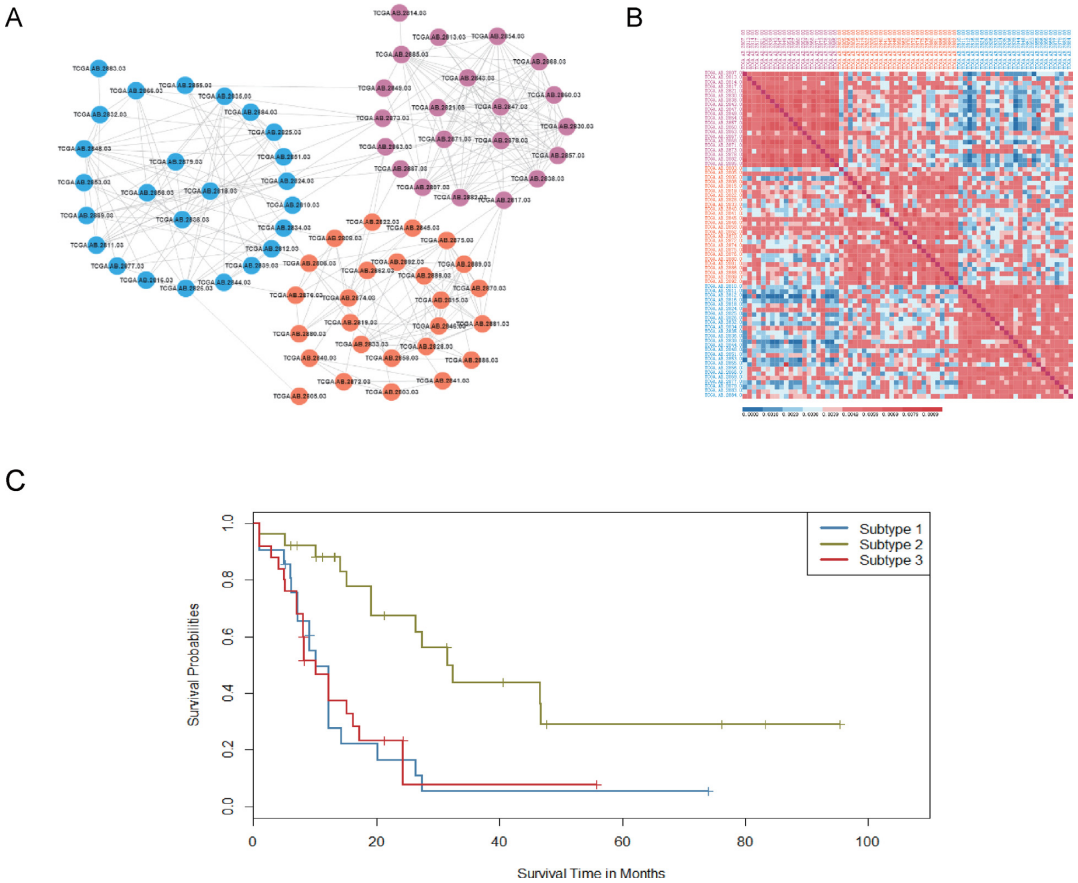

Figure 2. ICM results and survival curve for the three LAML subtypes that were identified from integrated data. (**A**) The similarity networks that were obtained for the patients with LAML. Nodes with the same color represent patient clusters. (**B**) A heatmap of patient data. Names (column names or row names) with the same color represent patient clusters. (**C**) A Kaplan–Meier survival curve for the three LAML subtypes (log rank P-value $= 2.9 \times 10^{-4}$).

Analysis Center [http://gdac.broadinstitute.org]): mRNA sequence, miRNA sequence and methylation data. These data were preloaded into the SNF web page as 'example data 3'. Using ICM, we subsequently integrated the three sets of omics data and clustered the samples into three subgroups.

As shown in Figure 2A and B, 21 patients were clustered as subtype 1, 26 patients were clustered as subtype 2 and 25 patients were clustered as subtype 3. The clustering result is valid and stable (connectivity = 38.9, Dunn index = 0.934, APN = 0.313). In the 'Result/Feature' subpage of the 'SNF' method, ICM ranked the features in each level of data using NMI and the top 1% of features were included in a heatmap to show the differences among the subtypes. According to the Kaplan–Meier method, the survival time of LAML patients between subtype 2 and others is associated with significant differences in survival curve (P = 0.00029) (Figure 2C). Besides, we separately use the mRNA sequence, miRNA sequence and methylation data to cluster the LAML patients. The survival time of LAML patients among subtypes is associated with no significant difference ($P = 0.15$ for mRNA sequence data alone, $P = 0.51$ for miRNA sequence data alone and $P = 0.73$ for methylation data alone). It indicates that integrated clustering is more efficient than clustering based on one data type alone.

DISCUSSION AND FUTURE DEVELOPMENTS

The ICM we have developed is an analysis web server that provides tools for the fusion, clustering and visualization of multi-dimensional biological data and knowledge.

The advantages of ICM include:

(i) **A wide range of potential users can access ICM.** In order to provide an analysis tool that can accommodate a variety of commonly analyzed complex objects, we designed ICM to not be limited to particular biomedical applications. Consequently, biologists, pharmacologists and clinicians should be able to use ICM in their research. For example, pharmacologists could use ICM to identify clusters of drugs based on structure, side effects, cell response, etc. for new indication discoveries, while clinicians could identify subtypes of patients based on available multi-dimensional clinical data.

(ii) **ICM provides three optional algorithms that have different characteristics.** The 'Concatenation' method is commonly used and is a simple and low computational-cost algorithm. Alternatively, the 'iCluster' method is based on a Gaussian latent variable model and can effectively discover potentially novel subtypes. This method employs a relatively high

computational complexity to analyze data with high dimension features. Finally, the 'SNF' method uses similarity network for samples in order to reduce the complexity of the computations performed. Thus, by using ICM, all three algorithms can be applied to an analysis of interest, while a comparison of the results from each algorithm is available as well.

(iii) **The analysis results can be visualized and modified for a publishable output.** On the ICM web site, users can preview their uploaded raw data and delete unsatisfactory features or sample data. ICM also provides a step-by-step guide for users to submit an analysis task. On the results page, a user can further modify the sample similarity network and heatmap to improve presentations of the results. Furthermore, the network presented by ICM can be saved as an image in .PNG and .SVG formats, and/or can be exported as a graphML file for analysis using Cytoscape.

Based on the potential applications of ICM to a wide range of research fields, it is anticipated that this web server will attract widespread interest among biomedical researchers. Moreover, we will continue to integrate new clustering algorithms into ICM in order to facilitate its use by scientists.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Nature Science Foundation of China [U1435222, 81273488]; National Key Technologies R & D Program for New Drugs [2012ZX09301-003]; Program of International S & T Cooperation [2014DFB30020]. Funding for open access charge: National Nature Science Foundation of China; National Key Technologies R & D Program for New Drugs; Program of International S & T Cooperation.
Conflict of interest statement. None declared.

REFERENCES

1. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., International Cancer Genome Consortium *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
2. Tomczak,K., Czerwinska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68–A77.
3. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
4. 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**, 98–110.
5. Kandoth,C., Schultz,N., Cherniack,A.D., Akbani,R., Liu,Y., Shen,H., Robertson,A.G., Pashtan,I., Shen,R., Cancer Genome Atlas Research Network *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.
6. 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.
7. 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.
8. 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.
9. Yuan,Y., Savage,R.S. and Markowetz,F. (2011) Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput. Biol.*, **7**, e1002227.
10. 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**, e35236.
11. Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Eng. J. Med.*, **372**, 793–795.
12. Guyon,I. and Elisseeff,A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.
13. Varshavsky,R., Gottlieb,A., Linial,M. and Horn,D. (2006). Novel unsupervised feature filtering of biological data. *Bioinformatics*, **22**, e507–e513.
14. Brock,G., Pihur,V., Datta,S. and Datta,S. (2011) clValid, an R package for cluster validation. *J. Stat. Softw.*, https://cran.r-project.org/web/packages/clValid/vignettes/clValid.pdf.
15. Handl,J., Knowles,J. and Kell,D.B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.
16. Dunn†,J.C. (1974). Well-separated clusters and optimal fuzzy partitions. *J. Cybern.*, **4**, 95–104.
17. Datta,S. and Datta,S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, **19**, 459–466.