BMC Bioinformatics

SOFTWARE Open Access

PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets

Dvir Netanely, Neta Stern, Itay Laufer and Ron Shamir*

Abstract

Background: Analysis of large genomic datasets along with their accompanying clinical information has shown great promise in cancer research over the last decade. Such datasets typically include thousands of samples, each measured by one or several high-throughput technologies ('omics') and annotated with extensive clinical information. While instrumental for fulfilling the promise of personalized medicine, the analysis and visualization of such large datasets is challenging and necessitates programming skills and familiarity with a large array of software tools to be used for the various steps of the analysis.

Results: We developed PROMO (Profiler of Multi-Omic data), a friendly, fully interactive stand-alone software for analyzing large genomic cancer datasets together with their associated clinical information. The tool provides an array of built-in methods and algorithms for importing, preprocessing, visualizing, clustering, clinical label enrichment testing, and survival analysis that can be performed on a single or multi-omic dataset. The tool can be used for quick exploration and stratification of tumor samples taken from patients into clinically significant molecular subtypes. Identification of prognostic biomarkers and generation of simple subtype classifiers are additional important features. We review PROMO's main features and demonstrate its analysis capabilities on a breast cancer cohort from TCGA.

Conclusions: PROMO provides a single integrated solution for swiftly performing a complete analysis of cancer genomic data for subtype discovery and biomarker identification without writing a single line of code, and can, therefore, make the analysis of these data much easier for cancer biologists and biomedical researchers. PROMO is freely available for download at http://acgt.cs.tau.ac.il/promo/.

Keywords: Cancer genomics, Personalized medicine, Cancer subtypes, Multi-omics, Cancer biomarkers, Multi-omic clustering, Gene expression analysis

Background

In recent years, a growing number of high-throughput genomic technologies have become available for biomedical research and are jointly providing high-resolution genomic data that fuel the revolution of personalized medicine [1, 2]. These technologies (collectively named omics) allow the simultaneous quantification of a large number of features at various biological levels. The features include gene expression (mRNA and miRNA abundance levels measured by microarrays or RNA-Seq), protein expression (measured by mass spectroscopy or reverse-phase protein arrays), DNA methylation (methylation arrays), copy number variation (SNP arrays), and others [3, 4]. The technologies vary broadly in the number of features they measure as well as in the distribution of measured values [5]. However, they can typically be summarized as a numeric matrix where columns represent samples and rows represent biological features (often correlating to genes). Bioinformatic analysis of such genomic matrices has been extensively used for identifying biologically distinct sample groups, and for revealing groups of correlated biological features [6, 7].

The number of tumor samples and measured features that are included in a typical cancer genomic dataset have grown dramatically in the last few years, owing to

* Correspondence: rshamir@tau.ac.il
Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

increasing resolution and reduced costs of array and sequencing technologies. Modern repositories comprise thousands of patient samples and many thousands of features. Investigation of such large datasets is computationally challenging as it requires robust software tools for supporting the analysis of both samples and features in high dimensional data [8].

In addition to genomic data, modern cancer datasets can include extensive medical information (labels) describing each sample, such as clinical properties or assignment to a predefined phenotype. These clinical labels make it possible to fuse genomic and clinical data in various ways in order to discover new insights based on feature-phenotype associations. Common clinical labels in cancer datasets include disease subtypes, pathological stages, survival and recurrence follow-up information, as well as response to treatment. Identification of genomic features that are correlated with significant clinical parameters (biomarkers) is expected to play a significant role in the field of personalized medicine, by which the status of multiple biomarkers may improve subtype diagnosis and guide therapeutic decisions [9, 10].

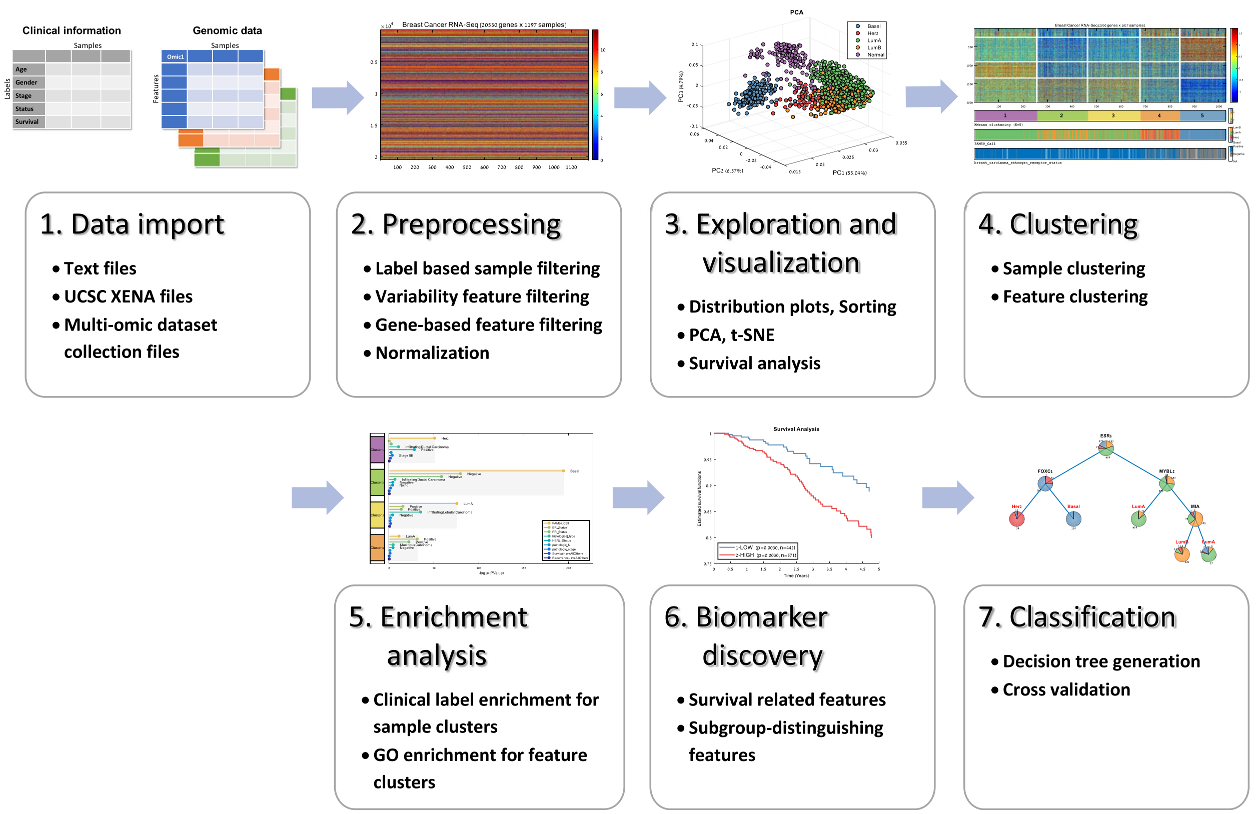
The Cancer Genome Atlas (TCGA) is an example of a revolutionary multi-label multi-omic genomic database [11]. It includes more than 11,000 samples from 33 types of cancer, where each sample was measured using multiple omic technologies and was described by dozens of clinical labels [12]. Many studies have already analyzed TCGA data, improving the subtyping of cancers and shedding light on the biological mechanisms underlying the development of various cancer types [13–15]. Such analyses are typically time-consuming, computationally challenging, and entail team effort, as they require applying a diverse array of methods, statistical tools, and algorithms, and often also require writing extensive computer code to perform and interweave the various steps of the analysis [16]. Hence, to effectively extract clinically meaningful insights from such multi-omic multi-label databases, specialized agile integrative tools are required.

To address this challenge, we developed **PROMO** (**PRO**filer of **M**ulti **O**mic data), a fully interactive software suite capable of quickly importing, preprocessing, visualizing, analyzing and reporting the results on cancer datasets in a seamless fashion, without writing a single line of computer code. PROMO includes an extensive array of bioinformatic methods for performing major common analysis types including exploration, visualization, identification of clinically significant disease subtypes, revealing co-regulated feature groups, biomarker discovery, simple classification and integrative multi-omic analysis. Table 1 presents an overview of the fundamental analysis types available in PROMO.

An early version of PROMO was developed as part of a study where we identified distinct prognostic subgroups in Luminal-A breast tumors based on expression and methylation data [17]. The analysis workflow in that project provides an example of the key steps in a typical application of PROMO (Fig. 1): Data are imported, filtered and preprocessed. Tumor samples are clustered into groups that are then assessed for clinical significance using survival analysis and statistical tests on the

Table 1 PROMO's main analysis types

	Analysis type	Biomedical goal	Relevant PROMO features
1	General exploration and visualization	Explore the genomic dataset vis-à-vis the clinical labels Prepare the dataset for downstream analysis, test its consistency and visualize its properties	• Variance-based feature filtering • Label-based sample filtering • Normalization • Sorting by sample label or mean expression • Visualizing data distribution • PCA, t-SNE
2	Focus on genes of interest	Explore the expression profiles of specific genes vis-à-vis multiple clinical labels Identify co-expressed genes	• Filter features based on gene symbols • Rank genes by correlation to a given gene symbol • Multi-label matrix visualization
3	Disease subtype identification	Look for clinically significant sample clusters	• Sample clustering • Label enrichment analysis • Survival analysis • Classification
4	Co-regulated feature group identification	Identify groups of similar features, characterize each group by function	• Feature clustering • GO Enrichment analysis
5	Biomarker discovery	Find features that distinguish among sample groups, correlate groups with survival and other clinical data	• Statistical tests for identifying differentially expressed genes • Biomarker-based survival analysis • Rank genes by survival prediction
6	Integrative multi-omic analysis	Stratify patients and identify coherent feature groups by integrating data from different omics	• Multi-omic sample clustering • Inter-omic feature correlation



1. Data import
- Text files
- UCSC XENA files
- Multi-omic dataset collection files

2. Preprocessing
- Label based sample filtering
- Variability feature filtering
- Gene-based feature filtering
- Normalization

3. Exploration and visualization
- Distribution plots, Sorting
- PCA, t-SNE
- Survival analysis

4. Clustering
- Sample clustering
- Feature clustering

5. Enrichment analysis
- Clinical label enrichment for sample clusters
- GO enrichment for feature clusters

6. Biomarker discovery
- Survival related features
- Subgroup-distinguishing features

7. Classification
- Decision tree generation
- Cross validation

Fig. 1 PROMO's subtype discovery workflow – From data import to subtype classifier. This figure outlines the complete workflow by which PROMO can be used for identifying and characterizing clinically distinct cancer subtypes: (1) Importing genomic data together with clinical information in one of several available formats. (2) Preprocessing the data and preparing it for downstream analysis. (3) Verifying the integrity of the data, characterizing its distribution and exploring dataset properties with respect to the available clinical labels. (4) Employing clustering algorithms partition both samples and features (genes) into groups. (5) Applying enrichment tests to identify clinically significant sample subtypes and groups of co-regulated genes. and to characterize their function. (6) Statistical tests identify features that distinguish between different sample subtypes as well as survival related features. (7) Decision tree classifiers can be generated for formulating a set of rules by which a new sample can be classified

clinical labels. Clustering of the genes followed by gene enrichment analysis associates sample clusters with active gene functions. The analysis is summarized visually in a genomic matrix clearly showing the identified sample clusters and their association to important clinical labels (Fig. 1, step 4), in addition to downstream analysis methods (Fig. 1, steps 5–7).

In this paper, we describe PROMO's main features and demonstrate its use in a study of a breast cancer cohort [14].

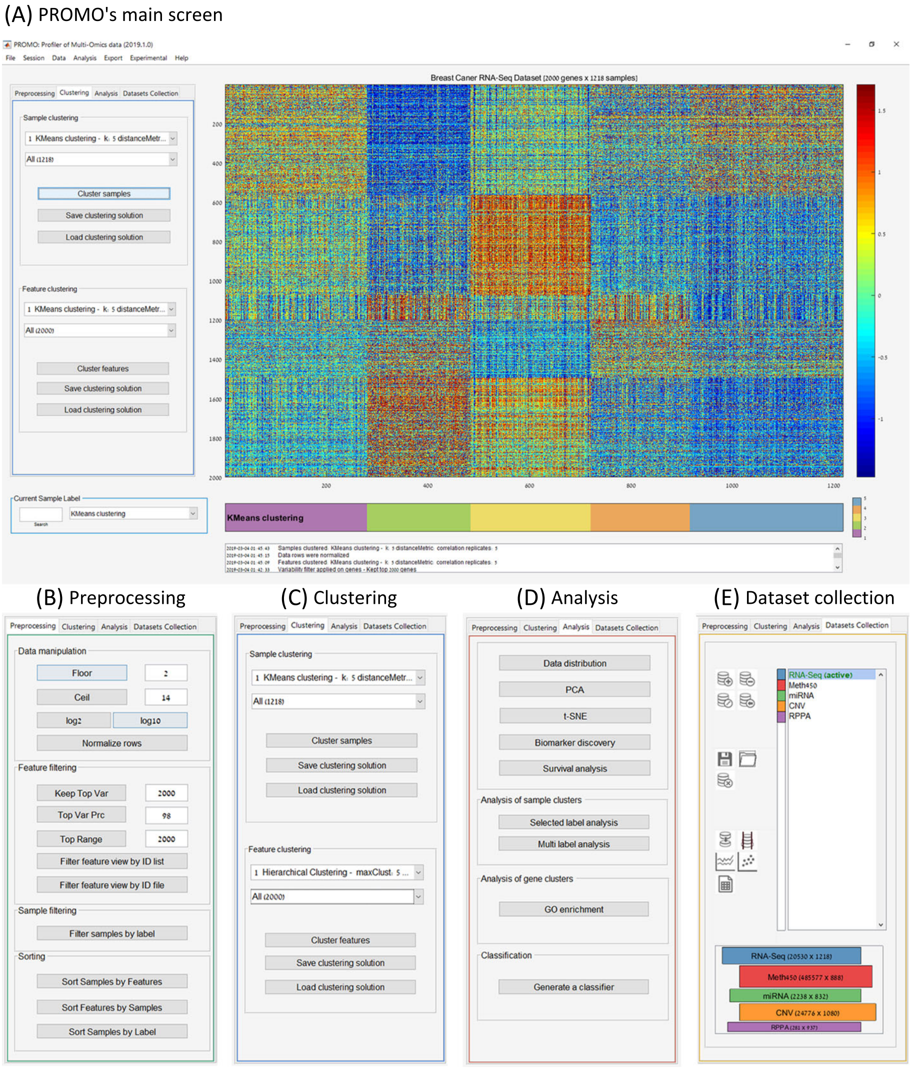
Results

We now describe PROMO's main features, organized by analysis steps. The described features can be accessed using PROMO's menus or graphical user interface (Fig. 2). The dataset used was TCGA's breast cancer gene expression profiles (1218 samples downloaded from UCSC's XENA website in May 2018). It is also available on the datasets page of PROMO's website.

Data import and preprocessing

In all analysis types, the first steps are to import the required data from local files into PROMO and prepare it for the analysis. PROMO enables the integration of data of different types and from multiple sources by importing genomic matrices, sample labels, and sample or gene partition files. Genomic matrices accompanied by complementary phenotypic information (clinical labels) can be loaded in the following formats: tabular text files, Gene Expression Omnibus (GEO) [18] series files (including direct download from within PROMO), UCSC's XENA [19, 20] file formats (available for many public datasets including all TCGA's data), and PROMO's DSC files. The latter are precompiled multi-omic datasets available at PROMO's dataset download page for selected TCGA cohorts. PROMO also allows separate loading of additional clinical labels and sample partition files to be used in the subtype discovery workflow.

After import, the loaded dataset can be 'cleaned' by filtering out samples based on clinical label values, and



```
2019-03-04 01:45:43   Samples clustered: KMeans clustering - k: 5 distanceMetric: correlation replicates: 5
2019-03-04 01:45:15   Data rows were normalized
2019-03-04 01:45:09   Features clustered: KMeans clustering - k: 5 distanceMetric: correlation replicates: 5
2019-03-04 01:42:33   Variability filter applied on genes - Kept top 2000 genes
```

Fig. 2 PROMO's graphical user interface. **a** PROMO's main screen. The genomic matrix is in the center with columns corresponding to samples and rows to features. Colors represent feature values according to the scale on the right. The colorful label bar beneath the matrix displays the currently selected sample label. Analysis steps are documented in the textbox on the bottom of the screen. Key commands are available on the tabbed panels on the left of the screen. **b** The *Preprocessing* panel allows filtering, normalization, and sorting of the genomic data. **c** Clustering the dataset's samples and features using various algorithms and distance functions is available through the *Clustering* panel. Resulting clustering solutions are aggregated for future review and filtering. **d** The *Analysis* panel provides access to several visualization and exploratory tools like PCA, t-SNE, survival analysis, biomarker discovery, GO enrichment and automatic classifier generation. **e** In the *Dataset Collection* panel, several genomic matrices can be assembled into a multi-omic dataset collection, and then analyzed together

also by removing certain features (e.g., removing low variability genes or keeping only specific genes). Additional available common preprocessing steps include flooring, ceiling, and row normalization.

Data exploration and visualization

Once a genomic matrix is loaded to PROMO, its properties can be explored with respect to any selected clinical label (Fig. 3a). The samples (columns) in the matrix can be reordered based on any clinical label or by their mean expression. Basic dataset properties like value distribution (Fig. 3b), clinical label distribution (Fig. 3c), and sample variation (Fig. 3d) can be studied and displayed graphically in various ways including PCA [21, 22] and t-SNE [23]. For ease of interpretation, all displays consistently use the same colors to represent the various sample subgroups.

Clustering and enrichment analyses

A major effort in promoting precision medicine is to identify disjoint groups of similar patients and characterize

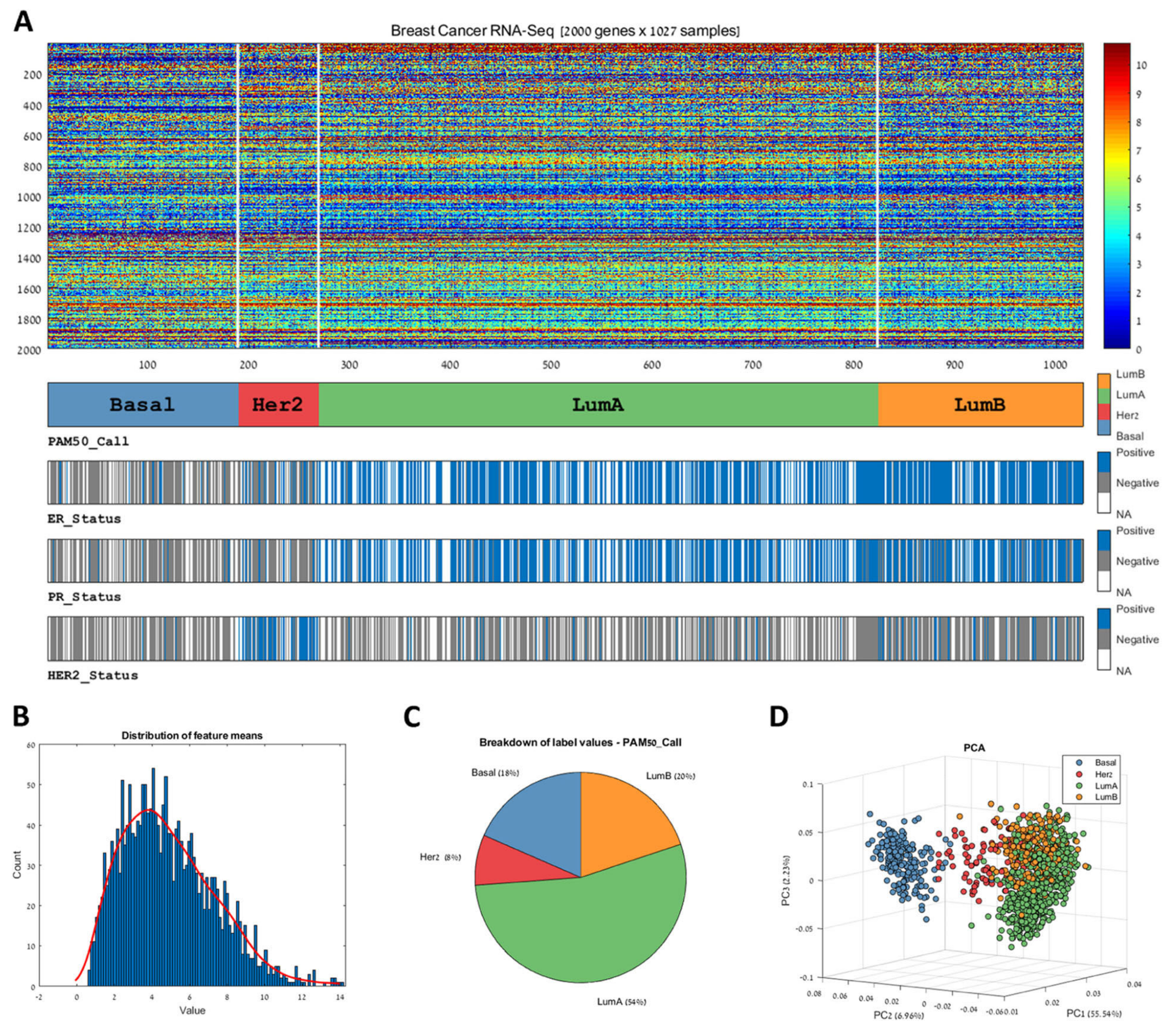


Fig. 3 Visualization of multi-label genomic data. PROMO provides a variety of methods for visualizing a genomic dataset together with its associated clinical information. **a** A multi-label expression matrix plot. The plot is composed of a heat-map representation of the genomic matrix and several label bars beneath it showing different clinical labels that the user interactively selected. The colors in each label bar show the label value of each sample according to the legend on the right. The label appears below the lower left corner of the bar. Here, breast cancer patient profiles were grouped according to their PAM50 category (shown in the top label bar). By observing the distribution of values in other bars, relations between the groups and the labels can be observed. For example, the ER, PR and HER2 status of most samples in the 'Basal' group are negative, while the HER2 status of most 'HER2' group is positive. **b** Data distribution and **c** Clinical label distribution can be explored and visualized separately, or in combination using plots such as **d** PCA and others. These figures show that the Basal tumor samples are mainly characterized by Negative ER, PR and HER2 labels (**a**) and markedly differ from all other subtypes in their gene expression pattern (**d**), in accordance with the literature [14]

each group using its distinct genomic profile, survival data, and clinical information. To reveal the similarities among patients, clustering is often performed on both samples and features [24]. Clustering the samples can reveal patient groups corresponding to disease subtypes [25] while clustering the features reveals groups of co-regulated genes [26]. PROMO provides various clustering algorithms such as K-means [27], hierarchical clustering [28], and Click [29] (PROMO's clustering panel is shown in Additional file 1: Figure S1). To explore the resulting clusters, the reordered matrix can be visualized in comparison to multiple sample labels (Fig. 4a).

After the genes have been clustered, the built-in Gene Ontology tool can help interpret the biological meaning of gene clusters using enrichment analysis (Fig. 4b) [30]. Likewise, the clinical labels on the samples can be used to statistically characterize each sample cluster. A comprehensive analysis can be applied to each sample cluster

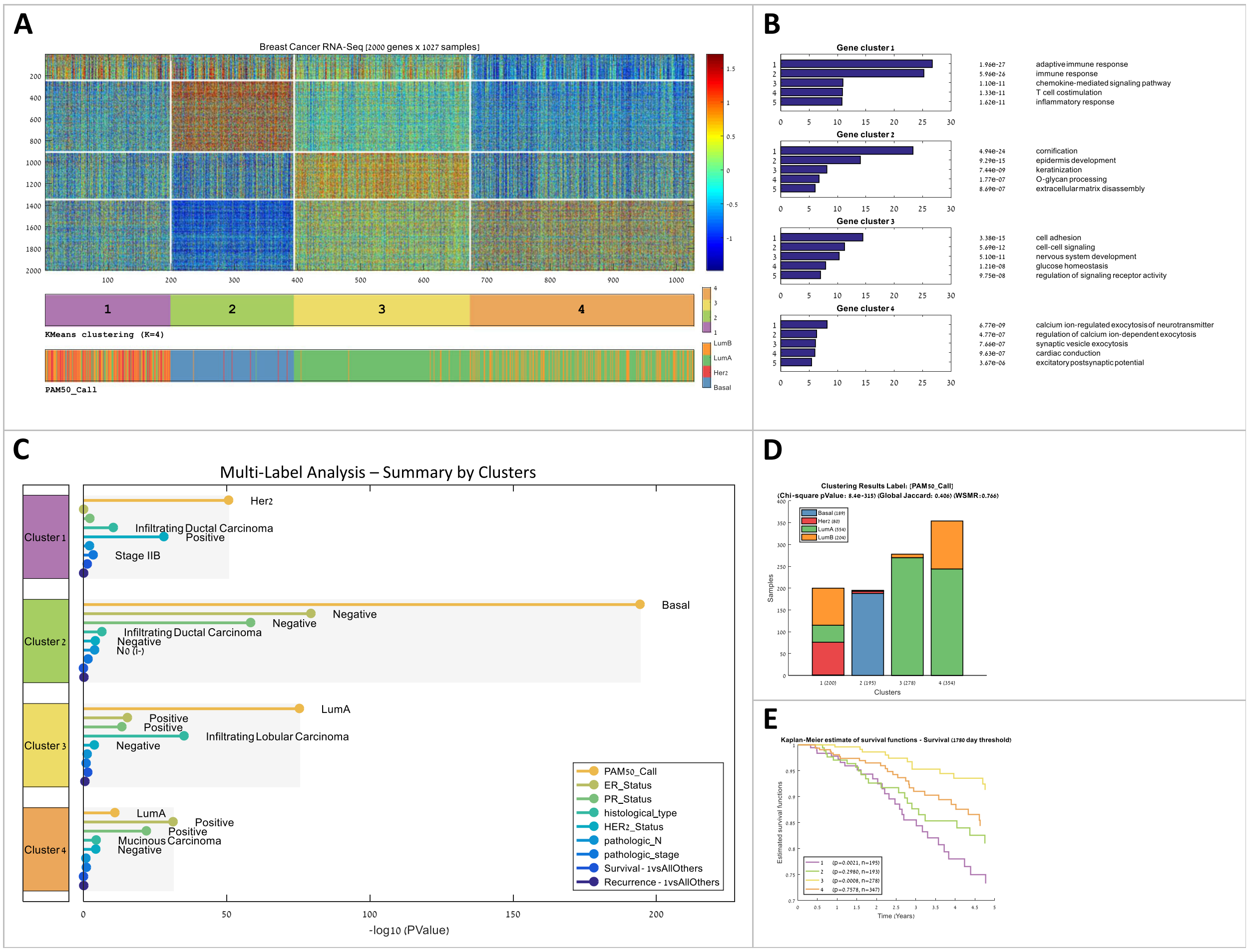


Fig. 4 Identification and characterization of cancer subtypes. Unsupervised analysis followed by enrichment analysis is performed on both samples and features for identifying clinically significant samples groups, and for biologically characterizing them based on the functions of co-expressed gene groups. **a** The RNA-Seq expression matrix of TCGA's breast cancer cohort after clustering both samples (columns) and genes (rows) into four clusters using the K-means algorithm. Clustering is based on the top 2000 variable genes. White lines separate clusters in each dimension. The bars below the matrix show selected sample labels (here: the clustering and PAM50). Matrix and bars were created using PROMO's multi-label matrix drawing. **b** Gene clusters were characterized using PROMO's gene ontology enrichment tool. The figure shows the five most significant GO terms for every gene cluster. **c**-**e** Sample clusters were characterized using the sample clinical labels: **c** PROMO's multi-label analysis tool automatically tests the clinical labels of different types (numeric, ordinal, categorical or survival) for enrichment on the sample clusters. FDR correction is performed over all clinical labels of the same type but separately for different types. The various **d** Sample clusters can also be characterized for a single label by showing its value distribution in each cluster and by calculating enrichment. **e** Survival functions for each cluster. The *p*-values are the significance of the separation of each cluster from the rest using the log-rank test

using all clinical labels available for the cohort (numeric, ordinal, categorical, or survival labels). The result is a characterization of each cluster, together with FDR corrected *p*-values [31, 32] in a unified report (Fig. 4c). Enrichment tests for the sample clusters can also be performed using any selected single clinical label (Fig. 4d). Finally, survival analysis performed on the sample clusters can test their prognostic value using Kaplan-Meier plots [33] and log-rank (Mantel–Haenszel) test [34](Fig. 4e). Taken together, PROMO's clustering and automatic multi-label enrichment analysis can quickly partition both samples and features into distinct groups and assess their biological meaning using the clinical labels.

Identification of distinguishing genes and features (biomarker discovery)

Having obtained patient subgroups of interest, either by sample clustering or using a predefined sample label, we may wish to identify distinguishing genes and features that differ significantly among sample groups. Such differentially expressed genes can shed light on the biological difference between sample clusters, and act as biomarkers for classifying a new sample to a sample class.

After selecting the label and the groups that will be compared, PROMO enables the application of various statistical tests for identifying genes that are differentially expressed among the groups. The *p*-values obtained by the tests can be used for gene sorting, filtering and for clustering the genes into up-regulated and down-regulated groups. PROMO's Gene Ontology enrichment analysis can be executed on the resulting gene groups for characterizing the function of up-regulated and down-regulated genes. FDR correction and fold-change based filtering are also supported. PROMO's biomarker discovery panel and an example of its output are shown in Additional file 1: Figure S2.

For detecting survival biomarkers, PROMO can rank all genes by their association to survival, based on Cox regression analysis [35]. In addition, the user can use the expression levels of selected genes to generate a new sample label (for example, HER2_Low and HER2_High). Kaplan-Meier plots can then be used to estimate the significance of survival differences between sample groups defined by the new label.

Lastly, PROMO can help in finding genes that are functionally related to a given gene of interest by ranking all genes based on their correlation to it. Altogether, the various techniques described here and implemented in PROMO can quickly identify genes that take part in the biological differences between sample groups and may serve as biomarkers for the selected label.

Automatic generation of a simple molecular classifier

After having partitioned the dataset samples, characterized the sample groups and their genes, and established the clinical relevance of the groups, PROMO can build an algorithm to classify a new sample into one of the groups. Such a classifier, especially if based on a small number of genes (rather than the thousands used to identify the subgroups) can serve as a significant step towards translating the analysis results into a diagnostic biomarker for clinical use.

Of the many possible classifier types, decision trees have the advantages of being easy to understand, highly interpretable biologically and easily visualized [36]. Furthermore, they allow for controlling the tradeoff between accuracy and simplicity. For predicting any selected sample label, PROMO can generate a simple decision tree with a single click (Fig. 5). The generated decision tree can be visualized graphically, specified textually, and saved to a Matlab file as a function. Automatic cross-validation and parameter optimization make it easy for the user to come up with a simple decision tree that may be in future subtype classification kits. It is also possible to generate a large number of random trees and rank the genes by the frequency of their appearance in

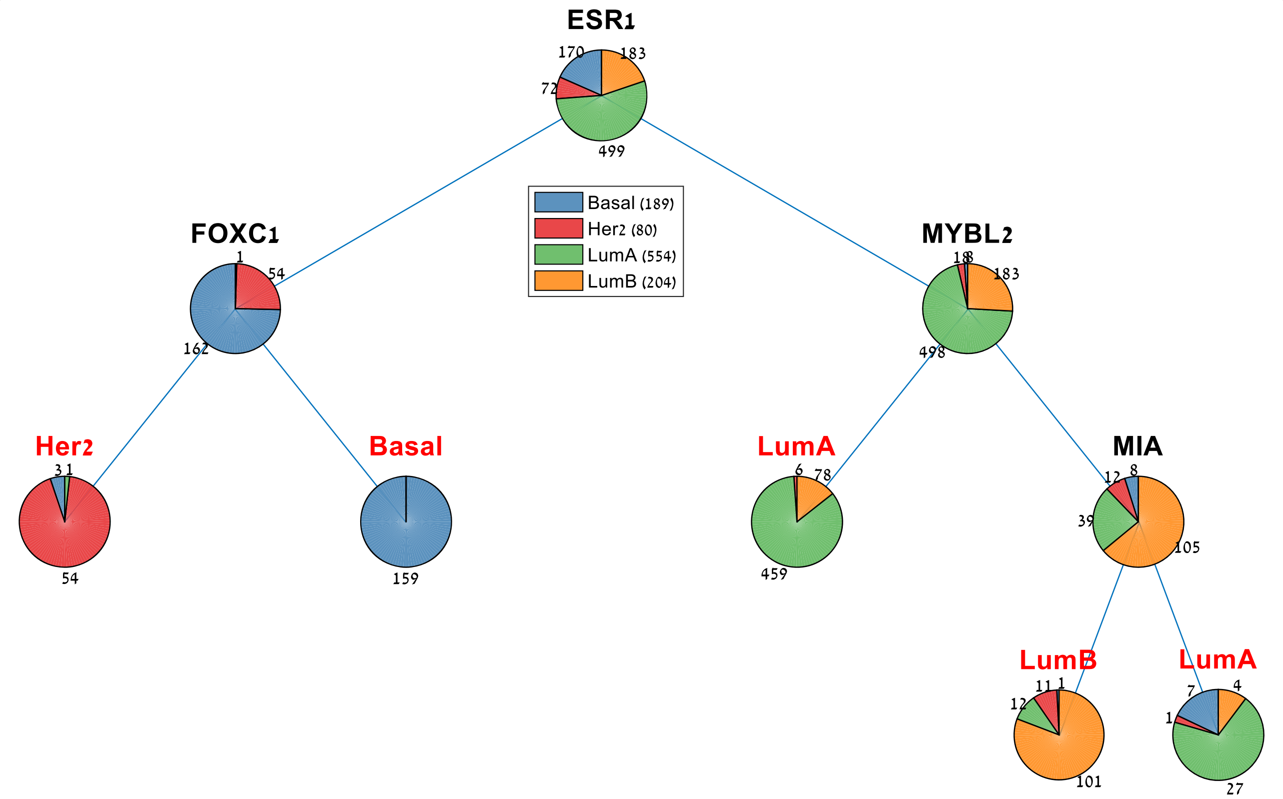


Fig. 5 Automatically generated decision tree for classifying breast tumors into the four PAM50 classes. PROMO can generate a cross-validated decision tree for any selected sample label using the currently loaded matrix as training data. In this figure, a four-gene molecular classifier for breast cancer subtypes is presented, showing a 7.77% loss on the training data, and a 15% averaged loss on 10-fold cross-validation

the trees, thus identifying informative features for subtype classification.

Integrative multi-omic analysis

In multi-omic datasets, each sample is characterized by several omic profiles (e.g., gene expression, methylation, copy number). Integrative analysis of multi-omic cancer datasets has the potential of revealing biological regulatory patterns that are missed in single omic analysis, and tools for performing such analyses are currently in great demand [37, 38].

PROMO provides several features for handling and analyzing multi-omic datasets. The profiles composing a multi-omic dataset can be imported from repositories into a 'Multi-Omic Dataset Collection' in PROMO (Fig. 2e). The user can navigate between the matrices, edit them independently, and select a subset of the datasets for downstream integrative analysis. Precompiled dataset collections for several TCGA cancer type cohorts are available on PROMO's download page.

After setting up a multi-omic collection, the "inter-omic correlation identification" feature helps to detect correlations between features in two selected omics. This feature allows the identification of correlations between features from different biological levels. For instance, anti-correlation between mRNA expression and DNA methylation levels can pinpoint biological regulation.

The "Multi-omic clustering" feature can be used to cluster the dataset samples based on several omic matrices simultaneously. To this end, PROMO provides implementations of the multi-omic algorithms SNF [39], NEMO [40], and Consensus Clustering [41] modified for multi-omic data. Additional file 1: Figure S4 demonstrates the application of a multi-omic clustering algorithm on three different omics of the TCGA's breast cancer cohort.

Discussion

Recent cancer projects such as TCGA [11], GDC [42], ICGC [43] as well as the GEO [18] database, provide the research community with a wealth of omic profiles and extensive clinical information on cancer patients [44]. Analysis of the data is challenging and requires advanced bioinformatics, statistical, and programming skills. A thorough analysis of these datasets - and larger ones expected in the future - by many researchers is crucial for improving cancer diagnosis and treatment.

PROMO aims to fill in a gap in available analysis tools for such large genomic and clinical cancer datasets. It is an interactive tool that is freely available and supports a rich collection of analysis methods and facilitates useful workflows for data exploration and visualization, cancer subtype identification, biomarker discovery and integrative multi-omic analysis. (See Table 2 for a list of the key features). PROMO's support for large sample size in addition to features like survival analysis and interrogation of the clinical data on sample clusters make it especially suitable for analyzing modern cancer datasets. While many of PROMO's features are also available in other tools (Table 3), PROMO is unique in its comprehensiveness, support for large sample dimension and the spectrum of tools it provides.

Our vision for PROMO is that it will be used as a one-stop-shop for mining clinically important insights from genomic datasets, quickly and without any need for programming skills. It accelerates the analysis process and makes it more accessible for non-computational cancer researchers. Within a single short session, the user can import a cancer dataset of interest, preprocess it, cluster its samples and features, test the sample clusters for significance using survival analysis and enrichment tests on the clinical labels, test the feature clusters for GO enrichment, identify subtype distinguishing features (biomarkers) using

Table 2 PROMO's key features

Category	Key Features
Data import	▪ Importing genomic data from tabular CSV files ▪ Importing UCSC's XENA genome matrix and phenotype files ▪ Importing GEO series files ▪ Adding clinical labels from file
Preprocessing	▪ Flooring, ceiling and row normalization ▪ Filtering of samples by clinical labels ▪ Filter features by range, variance, gene symbols or by an external list
Data exploration and visualization	▪ PCA, t-SNE ▪ Data distribution plots ▪ Survival Analysis (Kaplan Meier, Log rank) ▪ Multi-label expression matrix figures
Sorting	▪ Sorting samples and features based on genomic data ▪ Sorting samples based on clinical labels
Clustering	▪ Clustering both samples and features using K-means [27], hierarchical clustering [28], and Click [29] ▪ Browsing clustering history and zooming into specific clusters
Sample cluster analysis	▪ Automated multi-label enrichment test for detecting enrichment of clinical labels
Feature cluster analysis	▪ Gene ontology enrichment analysis
Biomarker discovery	▪ Applying statistical tests for detecting differentially expressed genes/features ▪ Filter results by FDR corrected *p*-value and fold change ▪ Rank genes based on survival prediction (COX regression)
Classifier generation	▪ Automatic generation of decision tree classifiers for selected sample labels
Integrative multi-omic analysis	▪ Assembly of dataset collection ▪ Multi-omic clustering using SNF [39], NEMO [40] or Consensus Clustering [41] ▪ Inter-omic correlation identification

Table 3 Comparison of the main functions provided by PROMO and by other tools

Function	PROMO	Expander [45]	XENA [20]	Perseus [46]	KnowEng [47]	O-Miner [48]
Precompiled datasets	*V*	*X*	*V*	*X*	*V*	*V*
Preprocessing	*V*	*V*	*X*	*V*	*X*	*V*
Data visualization	*V*	*V*	*V*	*V*	*V*	*V*
Sample clustering	*V*	*V*	*X*	*V*	*V*	*V*
Feature clustering	*V*	*V*	*V*	*V*	*X*	*V*
Sample clusters enrichment tests (clinical data)	*V*	*X*	*V*	*X*	*V*	*X*
Feature clusters enrichment tests	*V*	*V*	*X*	*V*	*V*	*V*
Survival analysis	*V*	*X*	*V*	*X*	*V*	*V*
Biomarker discovery	*V*	*V*	*X*	*V*	*X*	*V*
Automatic decision tree generation	*V*	*X*	*X*	*X*	*X*	*X*
Inter-omic correlation identification	*V*	*X*	*X*	*X*	*V*	*X*
Integrative multi-omic sample clustering	*V*	*X*	*X*	*X*	*X*	*X*

various statistical tests and export the results using various reports and figures. The simple classification capabilities in PROMO can automatically produce a decision tree classifier for any selected label, and thus act as a basis for a subtype diagnosis.

We intend to continue developing PROMO by adding features and supporting the tool's users. We hope that PROMO's comprehensiveness and ease of use will help cancer researchers make the best use of the accumulating cancer datasets to fulfill the promises of precision medicine.

Conclusions

PROMO is a powerful, user-friendly, stand-alone, publicly available tool for exploration, analysis, and interpretation of genomic cancer data together with clinical information.

Methods

Implementation

PROMO is a standalone Windows application that can support huge datasets and has a fast fully interactive graphical user interface. PROMO was written in MATLAB, and it runs over the freely available Matlab runtime environment, taking advantage of its strong computational engine and editable graphical outputs. PROMO is freely available for download at http://acgt.cs.tau.ac.il/promo/.

PROMO's main screen (Fig. 2a) includes several key graphic elements: A large heatmap representing the currently analyzed genomic matrix is located at the center of the screen (heatmap colors correspond to the matrix values as indicated by the color scale on the right). Beneath the heatmap, a color-bar displays the currently selected sample labels. The same sample label colors will consistently be used by PROMO in all displays. The user can scroll down the list of clinical labels and explore their distribution over the samples. The panel on the left provides access to common commands and parameters. A text log that documents the analysis steps appears at the bottom of the screen. Figures 2B-F show the various panels that can be directly opened from the tab menu on the left of the screen, providing quick access to PROMO's most useful features.

Availability and requirements

Project name: PROMO (Profiler of Multi-Omics data)
Project home page: http://acgt.cs.tau.ac.il/promo/
Operating system: Windows
Programming language: Matlab
Other requirements: Installation of Matlab runtime library R2019a (9.6)
License: GNU GPL 3.0
Any restrictions to use by non-academics: None

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-019-3142-5.

Additional file 1: Figure S1. Clustering Panel. **Figure S2.** Biomarker Discovery. **Table S1.** List of differentially expressed genes. **Figure S3.** Label Management Panel. **Figure S4.** Multi-omic sample clustering.

Abbreviations

BRCA: Breast Cancer; DSC: Dataset Collection; FDR: False Discovery Rate; GEO: Gene Expression Omnibus; GO: Gene Ontology; PCA: Principal Component Analysis; PROMO: Profiler of Multi-Omics data; TCGA: The Cancer Genome Atlas; t-SNE: t-distributed Stochastic Neighbor Embedding

Acknowledgements

Not applicable.

Declarations

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

Authors' contributions

DN designed and wrote the code. NS contributed to the design and implementation. IL contributed to the implementation. RS supervised the study. DN and RS wrote the manuscript. All authors have read and approved the final manuscript.

Funding

This study was supported in part by the Israel Science Foundation (ISF) as part of the ISF-NSFC joint program (grant 2193/15), ISF grant 1339/18, the Israel Cancer Association (donation of Avraham Rotstein), grant 2016694 from the United State - Israel Binational Science Foundation (BSF) and the United States National Science Foundation (NSF), and DIP German-Israeli Project co-operation grant.

The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Software and data are available at http://acgt.cs.tau.ac.il/promo/. Source code is available upon request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 29 May 2019 Accepted: 9 October 2019
Published online: 26 December 2019

References

1. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol. 2011;8:184–7.
2. Malod-Dognin N, Petschnigg J, Pržulj N. Precision medicine — a promising, yet challenging road lies ahead. Curr Opin Syst Biol. 2018;7:1–7.
3. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol. Cell. 2015;58:586–97.
4. MacConaill LE. Existing and emerging technologies for tumor genomic profiling. J Clin Oncol. 2013;31:1815–24.
5. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017;18:83.
6. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative methods for analyzing big data in precision medicine. Proteomics. 2016;16:741–58.
7. Roychowdhury S, Chinnaiyan AM. Translating cancer genomes and transcriptomes for precision oncology. CA Cancer J Clin. 2016;66:75–88.
8. Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: promises and challenges. Cancer Lett. 2013;340:284–95.
9. McDermott JE, Wang J, Mitchell H, Webb-Robertson B-J, Hafen R, Ramey J, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn. 2013;7:37–51.
10. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genet. 2015;8:33.
11. The Cancer Genome Atlas (TCGA) [Internet]. Available from: http://cancergenome.nih.gov/. Accessed 18 May 2018.
12. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemp. Oncol. 2015;19:A68–77.
13. Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20 Nature Publishing Group.
14. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.
15. The TCGA Legacy. Cell. Elsevier. 2018;173:281–2.
16. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2:84.
17. Netanely D, Avraham A, Ben-Baruch A, Evron E, Shamir R. Expression and methylation patterns partition luminal-a breast tumors into distinct prognostic subgroups. Breast Cancer Res. 2016;18:74.
18. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
19. Zhu J, Craft B, Goldman M, Cline M, Diekhans M, Haussler D. Using the UCSC Xena platform to integrate, visualize, and analyze your own data in the context of large external genomic datasets. Cancer Res. 2015;75(22 Suppl 2):Abstract nr B1-07.
20. Goldman M, Craft B, Hastie M, Repečka K, Kamath A, McDade F, et al. The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. BioRxiv. 2019:326470 Cold Spring Harbor Laboratory.
21. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip Rev Comput Stat. 2010;2:433–59.
22. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001;17:763–74.
23. García-Alonso CR, Pérez-Naranjo LM, Fernández-Caballero JC. Multiobjective evolutionary algorithms to identify highly autocorrelated areas: the case of spatial distribution in financially compromised farms. Ann Oper Res. 2014;219:187–202.
24. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med. 2008;38:283–93.
25. Saria S, Goldenberg A. Subtyping: What tt is and its role in precision medicine. IEEE Intell Syst. 2015;30:70–5.
26. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng. 2004;16:1370–86.
27. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28:129–37.
28. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci. 1998;95:14863–8.
29. Sharan R, Shamir R. CLICK: a clustering algorithm with applications to gene expression analysis. Proceedings. Int Conf Intell Syst Mol Biol. 2000;8:307–16.
30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9 Nature America Inc.
31. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57:289–300.
32. Bland JM, Altman DG. The logrank test. BMJ. 2004;328:1073.
33. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–81.
34. Horwitz RI. Statistical aspects of the analysis of data from retrospective studies of disease. J Chronic Dis. 1979;32:ii.
35. Cox DR. Regression models with life tables. J R Stat Soc Ser B. 1972;74:187–220.
36. Breiman L, Friedman J, Olshen R, Stone C. Classification And Regression Trees. Wadsworth: Chapman and Hall; 1984.
37. Vucic EA, Thu KL, Robison K, Rybaczyk LA, Chari R, Alvarez CE, et al. Translating cancer "omics" to improved outcomes. Genome Res. 2012;22:188–95.
38. Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. Front. Genet. Frontiers. 2017;8:84.
39. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.
40. Rappoport N, Shamir R. NEMO: Cancer subtyping by integration of partial multi-omic data. Bioinformatics. 2019;35:3348–56.
41. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118.
42. Genomic Data Commons Data Portal [Internet]. Available from: https://portal.gdc.cancer.gov/. Accessed 14 Feb 2018.
43. ICGC Data Portal [Internet]. Available from: https://dcc.icgc.org/. Accessed 5 Feb 2018.
44. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. Blood. 2017;130:453–9.
45. Ulitsky I, Maron-Katz A, Shavit S, Sagir D, Linhart C, Elkon R, et al. Expander: from expression microarrays to networks and functions. Nat Protoc. 2010;5:303–22 Nature Publishing Group.
46. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13:731–40.
47. Sinha S, Song J, Weinshilboum R, Jongeneel V, Han J. KnowEnG: a knowledge engine for genomics. J Am Med Inform Assoc. 2015;22:1115–9.
48. Sangaralingam A, Dayem Ullah AZ, Marzec J, Gadaleta E, Nagano A, Ross-Adams H, et al. "Multi-omic" data analysis using O-miner. Brief Bioinform. 2019;20:130–43.

Publisher's Note